MIL-HDBK-141

5 OCTOBER 1962

MILITARY STANDARDIZATION HANDBOOK

OPTICAL DESIGN

DEPARTMENT OF DEFENSE

UNITED STATES OF AMERICA

FSC 6650

DEFENSE SUPPLY AGENCY
WASHINGTON 25, D.C.

MIL-HDBK-141
Optical Design
5 October 1962

1. This handbook was developed by the Department of Defense with the assistance of a leading optical manufacturer. Major contributions were made by persons who, by virtue of experience in their particular fields, are recognized as qualified authorities on the subject of optical design.

2. This publication was approved on 5 October 1962 for printing and inclusion in the military standardization handbook series.

3. This document provides engineering personnel with an introduction to optical theory, and treats to an advanced level the fundamentals and principles of optical design. It is expected that wide distribution of the methods of design and computation presented in this handbook will result in a more efficient and accurate method of complying with military optical requirements.

4. To aid in maintaining the intended status of this handbook as a source of prevailing information, readers are encouraged to report any errors and suggestions for changes and additions to the Standardization Division, Defense Supply Agency, Washington 25, D. C.

This handbook contains copyright material.

CONTENTS

SECTION 1. INTRODUCTION 1-1
1.1 Scope 1-1
1.2 Definitions 1-2
1.3 Reference documents 1-5

SECTION 2. FUNDAMENTALS OF GEOMETRICAL OPTICS[1] 2-1
2.1 General 2-1
2.2 Law of refraction 2-1
2.3 Law of reflection 2-3
2.4 Total internal reflection 2-4
2.5 Index of refraction 2-4
2.6 Dispersion of light 2-5
2.7 Characteristics of optical glass 2-6

SECTION 3. CONSIDERATIONS OF PHYSICAL OPTICS[2] 3-1
3.1 Introduction 3-1
3.2 Physical nature of light 3-1
3.3 Interference between waves 3-4

SECTION 4. VISUAL OPTICS[3] 4-1
4.1 Introduction 4-1
4.2 Anatomy of the eye 4-1
4.3 Optical constants of the eye 4-3
4.4 Image formation and the retina 4-5
4.5 Seeing 4-10
4.6 Movement of the eyes 4-14
4.7 Binocular vision 4-16
4.8 Fatigue and ageing 4-18

SECTION 5. FUNDAMENTAL METHODS OF RAY TRACING[1] 5-1
5.1 General 5-1
5.2 Definitions and conventions 5-3
5.3 Basic ray trace procedure 5-5
5.4 Skew ray trace equations for spherical surfaces 5-5
5.5 Skew ray trace equations for aspheric surfaces 5-13
5.6 Meridional rays 5-21
5.7 Graphical ray tracing procedure 5-26
5.8 Differential ray tracing procedure 5-27
5.9 Paraxial rays 5-32
5.10 Graphical ray trace for paraxial rays 5-34
5.11 Different "orders" of optics 5-35

SECTION 6. FIRST ORDER OPTICS[1] 6-1
6.1 General 6-1
6.2 Numerical tracing of paraxial rays 6-1
6.3 Optical invariant 6-5
6.4 Linearity of paraxial ray tracing equations 6-8
6.5 Cardinal points of an optical system 6-10
6.6 Calculation of focal length from finite conjugate data 6-17

Ref 1, 2, 3 - See page vi for Author.

6.7 Systems of thin lenses in air 6-17
6.8 Optical systems involving mirrors 6-19
6.9 Differential changes in first order optics 6-23
6.10 Chromatic aberration 6-26
6.11 Entrance and exit pupils, the chief ray and vignetting 6-37

SECTION 7. SIMPLE THIN LENS OPTICAL SYSTEMS[1] 7-1

7.1 Introduction 7-1
7.2 Simple magnifier 7-1
7.3 Microscope 7-4
7.4 Telescope 7-8
7.5 Optical relay systems. Periscopes 7-8
7.6 Galilean telescope 7-11

SECTION 8. ABERRATION ANALYSIS AND THIRD ORDER THEORY[1] .. 8-1

8.1 Significance of ray trace data 8-1
8.2 Spot diagram 8-1
8.3 Meridional and skew fans 8-3
8.4 Use of third order theory in aberration analysis 8-3
8.5 Zero-degree image in D light 8-5
8.6 Imagery for an off-axis object point 8-9
8.7 Calculation of third order contributions 8-14
8.8 Afocal optical systems 8-15
8.9 Stop shift equations 8-16
8.10 Thin lens aberration theory 8-18

SECTION 9. METHOD OF LENS DESIGN[4] 9-1

9.1 Process of designing a lens system 9-1
9.2 Description and analysis of basic procedure 9-1
9.3 Summary of equations used in calculation of third order aberrations 9-11

SECTION 10. AN APPLICATION OF THE METHOD OF LENS DESIGN[4] . 10-1

10.1 Step one - selecting the lens type 10-1
10.2 Step two - first order thin lens solution 10-2
10.3 Step three - third order thin lens solution 10-13
10.4 Step four - thick lens first order and third order aberrations ... 10-17
10.5 Step five - tracing a few selected rays 10-19
10.6 Step six - readjusting third order aberrations 10-20
10.7 Evaluation of over-all performance 10-27
10.8 Summary 10-27

SECTION 11. TELESCOPE OBJECTIVES[4] 11-1

11.1 Introduction 11-1
11.2 Design procedure for a thin lens telescope objective 11-3
11.3 Design procedure for a thick lens telescope objective 11-6
11.4 Secondary spectrum of telescope objectives 11-18
11.5 Summary 11-25

SECTION 12. LENS RELAY SYSTEMS[4] 12-1

12.1 Introduction 12-1
12.2 The basic lens problem of a relay system 12-1
12.3 A visual system. Numerical example 12-1

Ref 1, 4 – See page vi for Author.

12.4 Secondary color in a relay system 12-2
12.5 Further details on design of doublets as relay lenses 12-2
12.6 Double relay systems 12-3
12.7 Summary 12-4

SECTION 13. MIRROR AND PRISM SYSTEMS 4 13-1

13.1 Introduction 13-1
13.2 Reflection 13-1
13.3 Location of image 13-4
13.4 Orientation of image 13-6
13.5 Image sphere 13-10
13.6 Reflection from two mirrors 13-13
13.7 Typical prism systems 13-15
13.8 Tunnel diagram 13-21
13.9 Aberrations introduced by prisms 13-23
13.10 Prism data sheets 13-25

SECTION 14. EYEPIECES 4 14-1

14.1 General principles 14-1
14.2 Method of description 14-1
14.3 Huygenian eyepiece 14-2
14.4 Ramsden eyepiece 14-4
14.5 Kellner eyepiece 14-6
14.6 Orthoscopic eyepiece 14-8
14.7 Symmetrical (Plössl) eyepiece 14-10
14.8 Berthele eyepiece 14-12
14.9 Erfle eyepiece 14-14
14.10 Modified Erfle eyepiece 14-16
14.11 Wild eyepiece 14-18
14.12 Summary 14-20

SECTION 15. COMPLETE TELESCOPE 4 15-1

15.1 Introduction 15-1
15.2 Design problem 15-1
15.3 Preliminary considerations 15-1
15.4 Design refinement 15-2
15.5 Completed design 15-3

SECTION 16. APPLICATIONS OF PHYSICAL OPTICS 2 16-1

16.1 Introduction 16-1
16.2 Fizeau interferoscope 16-3
16.3 Twyman-Green interferometer 16-5
16.4 Effect of monochromaticity on fringe contrast 16-7
16.5 Effect of pinhole size on contrast 16-8
16.6 Young's pinhole interferometer 16-9
16.7 Lloyd's interferometer 16-12
16.8 Fresnel coefficients for normal incidence 16-12
16.9 Interference with plane parallel plates and distant light sources . 16-14
16.10 Interference with plane parallel plates and nearby light sources . 16-16
16.11 Haidinger's interference fringes 16-17
16.12 Fizeau fringes 16-19
16.13 Newton's rings and Newton's fringes 16-19
16.14 Complex numbers 16-26
16.15 Transmittance of plane parallel plates 16-28
16.16 Reflectance from plane parallel plates 16-32
16.17 Multiple beam interference fringes from slightly inclined surfaces 16-34

Ref 4, 2 - See page vi for Author.

16.18 Measurements with monochromatic light 16-37
16.19 Method of channeled spectra 16-41
16.20 Interpretation of measurements with channeled spectra 16-41
16.21 Huygen's principle 16-45
16.22 Fraunhofer diffraction 16-47
16.23 Fraunhofer diffraction from a rectangular aperture 16-49
16.24 Fraunhofer diffraction from circular apertures 16-50
16.25 Diffraction from spherical wavefronts 16-52
16.26 Primary diffraction integrals with objectives having circular apertures 16-54
16.27 Resolution with circular apertures 16-56
16.28 Out-of-focus aberration 16-58

SECTION 17. OPTICAL MATERIAL [5] 17-1

17.1 Introduction 17-1
17.2 Refracting material characteristics 17-1
17.3 Refractivity and dispersion 17-3
17.4 Inclusions 17-4
17.5 Environmental characteristics 17-5
17.6 Refractive materials for specific wavelength ranges 17-5
17.7 Reflecting materials 17-8
17.8 Availability, cost, ease of working 17-10

SECTION 18. ATMOSPHERIC OPTICS [6] 18-1

18.1 Introduction 18-1
18.2 Extinction 18-1
18.3 Extinction and visual instruments 18-4
18.4 Extinction and photographic instruments 18-5
18.5 Seeing 18-6
18.6 Thermal effects 18-8
18.7 Atmospheric contaminants 18-9
18.8 Effect of atmospheric optics on instrument design 18-10

SECTION 19. OPTICS FOR MISSILE TRACKING [7] 19-1

19.1 Introduction 19-1
19.2 Refractive systems 19-2
19.3 Reflective systems 19-7
19.4 Catadioptric systems 19-10
19.5 Applied systems 19-14

SECTION 20. APPLICATIONS OF THIN FILM COATINGS [8] 20-1

20.1 Introduction 20-1
20.2 Manufacture of multilayer filters 20-14
20.3 Antireflection coatings 20-18
20.4 Reflectivity of multilayers with periodic structure 20-39
20.5 Long-wave pass filters 20-56
20.6 Short-wave pass filters 20-63
20.7 Beam splitters 20-64
20.8 Mirrors 20-68
20.9 Band pass filters 20-71
20.10 Fabry-Perot type filters (interference filters) 20-71
20.11 References for further syudy 20-91

Ref 5, 6, 7, 8 –See page vi for Author.

SECTION 21. COATING OF OPTICAL SURFACES [2] 21-1

21.1 Introduction 21-1
21.2 Definitions and principles 21-1
21.3 Zero reflectance from non-absorbing monolayers and substrates 21-27
21.4 Matrix methods 21-30
21.5 Quaternion methods 21-35
21.6 Monolayer coatings 21-41
21.7 Bilayer coatings 21-49
21.8 Trilayers 21-64
21.9 Quadrilayers 21-66
21.10 Quarter-wave multilayers 21-67
21.11 Materials and texts 21-77

SECTION 22. INFRARED OPTICAL DESIGN [5] 22-1

22.1 Introduction 22-1
22.2 Infrared optical material 22-1
22.3 Environmental requirements 22-2
22.4 Operational requirements 22-2
22.5 Near infrared region 22-3
22.6 Intermediate and far infrared region 22-5
22.7 Summary and conclusion 22-12

SECTION 23. MICROSCOPE OPTICS [9] 23-1

23.1 Introduction 23-1
23.2 Characteristics 23-1
23.3 Components of a compound microscope 23-3
23.4 Darkfield microscopy 23-11
23.5 Ultramicroscopy 23-15
23.6 Phase microscopy 23-16
23.7 Interference microscopy 23-19
23.8 Polarizing microscopy 23-23
23.9 Fluorescence microscopes 23-23
23.10 Stereoscopic microscope 23-24
23.11 Petrographic microscope 23-24

SECTION 24. DESIGN PHASE OPTICAL TESTS [2] 24-1

24.1 Introduction 24-1
24.2 Calculation of Seidel aberrations 24-2
24.3 Spot diagram 24-3
24.4 Phase front calculations 24-4

SECTION 25. PRODUCTION PHASE OPTICAL TESTS [2] 25-1

25.1 Introduction 25-1
25.2 Focal length 25-1
25.3 Longitudinal spherical aberration 25-3
25.4 Coma 25-4
25.5 Astigmatism and curvature field 25-4
25.6 Distortion 25-5
25.7 Auxiliary optical measurements 25-5
25.8 Optical devices, testing systems and procedures 25-8
25.9 Ronchi test 25-20
25.10 Foucault test 25-24
25.11 Star test 25-29

Ref 2, 5, 9 - See page vi for Author.

SECTION 26. EVALUATION PHASE OPTICAL TESTS [2] 26-1
26.1 Resolving power tests 26-1
26.2 General discussion of sine-wave testing 26-10
26.3 Sine-wave testing with sine-wave targets 26-12
26.4 Sine-wave testing with square-wave targets 26-17

APPENDIX 27-1

INDEX 28-1

REF	AUTHOR
1.	Dr. Robert E. Hopkins, University of Rochester Dr. Richard Hanau, University of Kentucky
2.	Dr. Harold Osterberg, American Optical Company
3.	Dr. Oscar W. Richards, American Optical Company
4.	Dr. Robert E. Hopkins, University of Rochester
5.	Mr. A. J. Kavanagh, American Optical Company
6.	Dr. Ralph Wight, Photronics Corporation
7.	Dr. Seymour Rosin, Scanoptic, Incorporated
8.	Dr. Philip Baumeister, University of Rochester
9.	Mr. Alva Bennett, American Optical Company

1 INTRODUCTION

1.1 SCOPE

In 1952 the Ordnance Corps published ORDM 2-1, Design of Fire Control Optics. The purpose of that publication was to make available to engineering and design personnel all pertinent optical data that had been accumulated by Frankford Arsenal. In the meantime, the rapidly increasing application of optical features in the design of military systems, and the accelerated rate of over-all technical advancement in the optical field bypassed ORDM 2-1 to such an extent that, in 1958, a tri-service project was initiated to gather and present, in a single volume, up-to-date engineering information, formulas, and calculations currently applicable in the design of individual optical elements and complete optical systems. Military handbook MIL-HDBK-141 is the result of that project.

This Department of Defense handbook was developed by a leading optical manufacturer under Department of the Army Contract DA-36-038-ORD-20590. Major contributions were made by a group of recognized authorities in the field of optical design. All work was performed under the guidance of Frankford Arsenal.

Although many excellent reference works at the college and advanced-design level are available, there is a lack of transition among them from one subject to another. To provide this needed transitional feature, MIL-HDBK-141 presents as nearly as possible the full range of subjects encountered in the field of optical design, including sections covering fundamentals, principles of design, and design data.

The first seven sections serve mainly to acquaint the reader with the basic concepts of optics, and to introduce the mathematical notation employed in later sections. These initial sections require that the reader have a working knowledge of analytical geometry, differential and integral calculus, and physics.

The sections on principles of design introduce typical design considerations encountered in basic types of optical systems. Included are discussions on sysem aberrations and their computation and correction. The computing schemes described should enable the designer to work efficiently and accurately.

The remaining sections of the handbook apply to various commonly used components and combinations, discussions of problems and solutions in special design areas, and data on general topics related to problems of optical design and manufacture.

1.2 DEFINITIONS.

1.2.1 Symbols and Notations. The following symbols are used in this handbook. Table I contains the English alphabet notation, Table II, the Greek alphabet, and Table III, the mathematical symbols.

TABLE I

Symbol	Usage
A	Area, points, linear dimension.
$\overline{A}$	Aperture area.
a	Real number, points, mirror aperture, amplitude of a wave. Special (Geometrical optics): 3rd order chromatic aberration.
B	Points, 3rd order surface contribution for spherical aberration.
b	Real number, coefficient of a power series. Special (Geom. optics): 3rd order chromatic aberration.
C	Points. As a subscript denotes red light using hydrogen line. Special (Geom. optics): 3rd order surface contribution for astigmatism.
c	Points, constant velocity for all electromagnetic waves in a vacuum, curvature; positive if the center of curvature is to the right of a surface.
D	Lens diameter. As a subscript denotes yellow light using sodium line. Special (Geom. optics): distortion.
d	Thickness, pupil diameter in mm., as a subscript denotes yellow light using helium line. Special (Geom. optics): distance or distance along a ray (not along optical axis).
E	Electric vector, 3rd order contribution for distortion.
F	Principal focal point. As a subscript denotes blue light of hydrogen line. Total flux radiated by a surface.
f	Focal length of a lens; positive if the first principal point is to the right of the first principal focus. A function related to the phase changes on reflection at the reflecting coated surfaces.
f'	Focal length of a lens; positive if the second principal focal point is to the right of the second principal point.
H	Magnetic vector.
h	Diameter of a mirror, fringe width.
I	Angle of incidence; positive if the ray can be made coincident with the normal to the surface by rotating the ray in a clockwise direction by an angle less than 90°.
I'	Angle of refraction; positive if the ray can be made coincident with the normal to the surface by rotating the ray in a clockwise direction by an angle less than 90°.
I_c	Critical angle.
i	Imaginary number, paraxial angle of incidence.
i'	Paraxial angle of refraction.
$J_n(x)$	Bessel function of order n.
K	Absorption constant, constant of proportionality, optical constant, optical direction cosine, Ratio of the energy density at the diffraction head when objective is out-of-focus by an amount to the energy density at the diffraction head when objective is in focus.
k	Image surface, $2\pi/\lambda$.
L.	Distance, optical direction cosine.
l	Path length through the particular medium.
M	Magnification ratio, optical direction cosine, unit normal vector.
MP	Magnifying power.
m	Lateral magnification.
N	Number of inter-reflections, nodal point of a lens.
n	Index of refraction, optical constant of an homogeneous isotropic film, n^{th} order of terms.
O	Origin, object surface.

TABLE I (Cont.)

Symbol	Usage
o	As subscript pertains to object.
P	Object point, principal point of a lens. Special (Geom. optics): Petzval contribution.
P'	Image point.
$\tilde{P}$	Partial dispersion ratio.
PD	Interpupillary distance.
Q	Incident unit ray vector, quaternion, ratio.
Q'	Reflected unit ray vector.
q	Scalar coefficient.
R	Radius, reflectance, resolving power in seconds of arc.
r	Radius.
S	Object conjugate of a lens, surface of a lens, time-averaged Poynting vector.
S'	Image conjugate of a lens.
T	Internal transmittance, time-averaged energy transmittance, period.
t	Thickness measured along optical axis, Special: (Physical optics): time.
U	Angle between meridional ray and optical axis, vector.
u	Angle between paraxial ray and optical axis, polar coordinate.
V	Distance, optical path, vector, wavefront.

Symbol	Usage
v	Velocity of light in vacuum, size of field of view.
W	Energy density or energy flux.
w	Optical half-width of the Fabry-Perot fringes.
X, Y	Rectangular coordinate system of the Z plane, with subscripts they denote the position coordinate of the ray intercepts on the subscript surface.
X_ν	Radii of the dark fringes.
X_μ	Radii of the bright fringes.
x	Distance along X-coordinate.
Y	Radius of entrance pupil.
$\overline{Y}$	Height of chief ray.
Y_ν	Admittance when electric vector is perpendicular to the plane of incidence in the νth layer.
$\overline{y}$	Object height, height of oblique paraxial ray.
y_ν	Admittance when the magnetic vector is perpendicular to the plane of incidence in the νth layer.
Z	The abscissa of the rectangular coordinate system used. In general the axis of propagation or optical axis; with subscript, denotes a position coordinate of the ray intercept on the subscript surface. Complex number, sag.
z	Distance along Z axis.

TABLE II

Symbol	Usage
α	Absorption coefficient, angle, angular magnification, direction cosine with respect to X axis.
β	Absorption coefficient, angle, direction cosine with respect to Y axis.
γ	Constant, direction cosine with respect to Z axis.
Δ	Total phase difference, increment of change.
δ	Angle of deviation, phase difference.
ϵ	Dielectric constant.
ζ	Abscissa.
η	Ordinate.
κ	Extinction coefficient.
θ	Angular limit of resolution, angular measurement.
λ	Wavelength.
μ	Magnetic permeability.
ν	Abbe constant, extinction coefficient, frequency of vibration, integer.
ρ	Amplitude reflectance.
σ	Electric conductivity, phase change, unit vector.
τ	Amplitude transmittance.
ϕ	Angle, phase angle, power of a thin lens.
Φ	Optical invariant.
ω	Angular velocity, angle.

TABLE III

Symbol	Usage
$\pm$	Plus or minus.
$=$	Equal to.
$\equiv$	Identity, defined as.
$\approx$	Nearly equal to.
$\sim$	Similar to, special designator when used to overline a capital letter.
$\rightarrow$	Approaches (from left hand side).
$\leftarrow$	Approaches (from right hand side).
$>$	Greater than.
$<$	Less than.
$\leqq$	Less than or equal to.
$\geqq$	Greater than or equal to.
$^\circ$	Degree.
$\therefore$	Therefore.
$(\ )$	Parentheses; multiplication operator.
$*$	Transverse chromatic aberration for some oblique ray displaced from the ray passing through $y_1=0$.
$\sqrt{\ }$	Square root.
$\sqrt[n]{\ }$	nth root.
Σ	Summation operator.
$\sum$	Sigma-summation operator.
∞	Infinity.
$[\ ]$	Brackets; multiplication or matrix operators.
∂	Denotes partial differentiation.
$\int$	Integration operator.
$\displaystyle\int$	Integration operator.
π	Pi = 3.1416. π radian = 180°.
e	Base of Naperian or natural logarithm = 2.71828.
$\prod$	Quaternion summation operator.

1.2.2 Terms. In general, the terms used in this handbook conform to Military Standard No. 1241, Optical Terms and Definitions; where special terms are used, the definitions are given in the text. An alphabetical index is provided at the end of the volume for easy reference to these definitions.

1.3 REFERENCE DOCUMENTS.

1.3.1 The following government publications are used in direct reference or provide related information valuable in the general field of optical design:

JAN-G-174 Optical Glass
MIL-STD-12 Abbreviations for Use on Drawings
MIL-STD-34 General Requirements for the Preparation of Drawings for Optical Elements and Optical Systems
MIL-STD-106 Mathematical Symbols
MIL-STD-150 Photographic Lenses
MIL-STD-1241 Optical Terms and Definitions

1.3.2 The following commercial publications are used in direct reference or provide related information valuable in the general field of optical design:

Ballard, S. S., McCarthy, K. A., Wolfe, W. L., State-of-the-Art Report: Optical Materials for Infrared Instrumentation, (Report No. 2389-11-S: I. R. I. A, Univ. of Michigan, 1959).
Bennett, A. H., Jupnik, H., Osterberg, H. and Richards, O. W., Phase Microscopy, (John Wiley and Sons, 1951).
Born and Wolf, Principles of Optics, (Pergamon Press, 1959).
Committee on Colorimetry, The Science of Color, (Thomas Crowell Co., 1954).
Conrady, Applied Optics and Optical Design, Parts 1 and 2 (Dover Publications Inc., 1960).
Drude, Theory of Optics, (Dover Publications Inc., 1960).
Hardy and Perrin, The Principles of Optics, (McGraw - Hill, 1932).
Holland, L., Vacuum Deposition of Thin Films, (John Wiley and Sons, 1956).
International Lighting Vocabulary Vol. I. (CIE. - 1.1. - 1957).
Jacobs, Fundamentals of Optical Engineering, (McGraw-Hill, 1943).
Jenkins and White, Fundamentals of Optics, (McGraw-Hill, 1957).
Johnson, B. K., Optics and Optical Instruments, (Dover Publications Inc., 1960).
Journal, Optical Society of America.
Linfoot, Recent Advances in Optics, (Oxford, 1955).
Martin, Technical Optics, (Pitman, 1948).
National Bureau of Standards, Circular No. 526, Optical Image Evaluation, (1954).
Optical Industry Directory, (Optical Publishing Co., 1961).
Sawyer, Experimental Spectroscopy, (Prentice-Hall, 1951).
Searle, Experimental Optics, (Cambridge Univ. Press, 1926).
Sears, F. W., Optics, (Addison-Wesley Press, Inc., 1949).
Strong, Concepts of Classical Optics, (Freeman, 1958).
Strong, Procedures in Experimental Physics, (Prentice-Hall, 1953).
Taylor, The Adjustment and Testing of Telescope Objectives, (Grubb, Parsons and Co., 1946).
Twyman, Prism and Lens Making, (Hilger, 1957).
Wagner, Experimental Optics., (John Wiley and Sons, 1929).

2 FUNDAMENTALS OF GEOMETRICAL OPTICS[1]

2.1 GENERAL

2.1.1 Geometrical optics. The term geometrical optics is applied to that branch of physics which deals with the propagation of light in terms of rays. These rays are considered as straight lines in homogeneous media. Geometrical optics, however, does not include some of the wave aspects of light propagation and hence does not take into account interference or diffraction effects. It is the starting point of the design of all optical systems; often it is the end point. It offers a means of progressing from graphical representations to numerical methods of analysis, and of arriving at solutions which in most cases are sufficiently accurate. One purpose of this text is to describe the laws and principles of geometrical optics and to show their application to the design of optical elements and systems.

2.1.2 Wave surfaces and rays. A basic problem in the design of optical systems is the calculation of wave surfaces as they progress through the various optical media. In geometrical optics this calculation is approximated by considering a relatively small number of rays, and then tracing these rays through the system. The actual passage of the rays is computed using analytic geometry procedures and two simple laws, the law of reflection and the law of refraction.

2.1.3 Direction of rays. The rays are perpendicular to the wave surfaces if the radiation is passing through a medium which is optically isotropic. The position of a wave surface (often called a wavefront) with respect to a point source may be determined at any time by the following procedure. From the point source equal optical path lengths are laid off along the rays. The surface that passes through these end points and is normal to the rays is a wavefront. (The optical path length corresponding to a physical path length is the product of the physical path length and the index of refraction.) In birefringent material the ray directions are not necessarily normal to the wave surfaces. The path of a ray of light traveling in a homogeneous medium is a straight line. When the ray is incident upon a surface separating two optically different media, it is reflected and refracted. This usually results in an abrupt change in the direction of the ray.

2.1.4 Angles of incidence, reflection, and refraction. If a normal is erected to the surface separating two media at the point where the ray is incident, the angles which the normal makes with the incident, refracted, and reflected rays are termed, respectively, the angles of incidence, refraction, and reflection. The laws of refraction and reflection, which state the relations existing between these angles, are two of the fundamental laws upon which optical design is based. The third law, mentioned above, states that a ray in a homogeneous medium travels in a straight line.

2.2 THE LAW OF REFRACTION

2.2.1 Diagram for refraction. Figure 2.1 shows a ray of light refracted at an interface between two different homogeneous materials characterized by n_0 and n_1, which are the respective indices of refraction of of the materials. The interface is shown as a straight line representing the intersection of a plane surface with the plane of the paper. This is a special case of the general situation in which the interface is a curved surface. In addition to the refracted ray, shown in Figure 2.1, in general there will also be a reflected ray. This has been omitted in the figure only for the purpose of clarification. For most cases where refraction is the aim, the reflected rays account for less than 10% of the incident energy. Section 21.2 will discuss the calculation of the reflected energy.

2.2.2 Sign convention. The following sign convention will be used for the angles of incidence, refraction, and reflection. If the ray must be rotated clockwise through the acute angle to bring it into coincidence with the normal to the surface, the angle is called positive. The angles I and I' in Figure 2.1 are both positive.

2.2.3 Statement of the law of refraction. The law of refraction is stated in two parts:

(1) The incident ray, the refracted ray, and the normal to the surface all lie in a single plane.

(2) The sines of the angles of incidence and refraction are related by the equation

$$n_0 \sin I = n_1 \sin I' . \quad (1)$$

2.2.4 Vector form of the law of refraction.

2.2.4.1 In solving many three dimensional refraction problems it is convenient to express the law of refraction in vector form. This is accomplished by describing the incident ray direction by a vector of unit

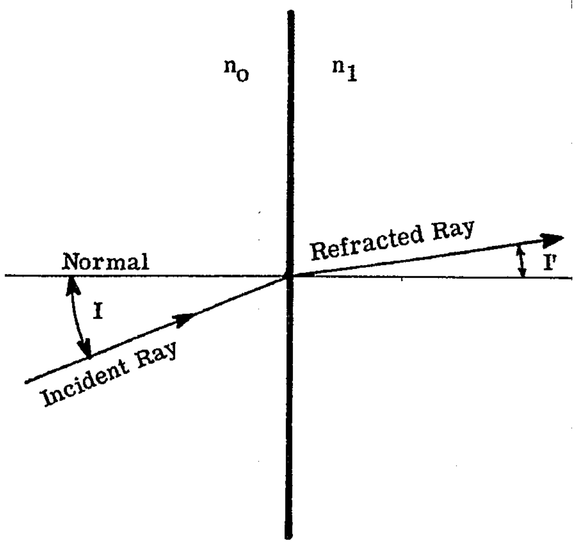


Figure 2.1 - Illustration of refraction.

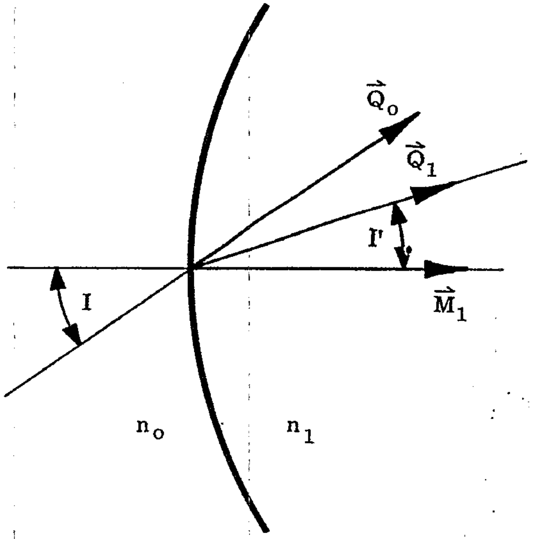


Figure 2.2 - Unit vectors for ray directions.

length $\vec{Q}_o$, the refracted ray by a unit vector $\vec{Q}_1$, and the normal by a unit vector $\vec{M}_1$. Figure 2.2 shows the relationship between these unit vectors and the ray directions. The vector $\vec{M}_1$ lies along the normal in the direction incident medium to refractive medium.

2.2.4.2 The vector product (cross product) of the two vectors $\vec{Q}_o$ and $\vec{M}_1$ is a vector of magnitude

$$\vec{Q}_o \times \vec{M}_1 = |Q_o| \; |M_1| \sin I = \sin I$$

because the angle between these vectors is I and they are each of unit length. The vector whose magnitude is sin I is perpendicular to the plane containing angle I (the plane of Figure 2.2), and directed perpendicularly into the plane of the paper. Similarly, $\vec{Q}_1 \times \vec{M}_1 = \sin I'$, and this is a vector parallel to $\vec{Q}_o \times \vec{M}_1$, because the refracted ray lies in the plane determined by the normal and the incident ray.

2.2.4.3 We have established the parallelism of the two vectors whose magnitudes are sin I and sin I' . By Equation (1) their magnitudes are in the ratio of the indices. Hence

$$\frac{\sin I}{\sin I'} = \frac{\vec{Q}_o \times \vec{M}_1}{\vec{Q}_1 \times \vec{M}_1} = \frac{n_1}{n_o} ,$$

and the vector form of the law of refraction may be written as

$$n_o (\vec{Q}_o \times \vec{M}_1) = n_1 (\vec{Q}_1 \times \vec{M}_1) . \qquad (2)$$

Equation (2) indicates, as all vector equations do, that the vector given by the left hand side equals in magnitude and direction the vector given by the right hand side.

2.2.4.4 Equation (2) can be written in another form by absorbing the scalar quantities n_o and n_1. Replacing the two vectors $n_o \vec{Q}_o$ and $n_1 \vec{Q}_1$ by $\vec{S}_o$ and $\vec{S}_1$, respectively, we have

$$\vec{S}_o \times \vec{M}_1 = \vec{S}_1 \times \vec{M}_1 ,$$

and

$$(\vec{S}_1 - \vec{S}_o) \times \vec{M}_1 = 0 .$$

Since neither $\vec{M}_1$ nor $(\vec{S}_1 - \vec{S}_o)$ is zero, these two vectors must be parallel or anti-parallel. Therefore we can define a quantity Γ (sometimes called the astigmatic constant) by writing

$$\vec{S}_1 - \vec{S}_o = \Gamma \vec{M}_1 . \tag{3}$$

2.2.4.5 Having found the direction of $(\vec{S}_1 - \vec{S}_o)$, we now want to determine its magnitude, Γ. From the definitions of $\vec{S}_o$ and $\vec{S}_1$, and because $\vec{Q}_o$ and $\vec{Q}_1$ are unit length, $\vec{S}_o$ and $\vec{S}_1$ are two vectors of length n_o and n_1, in the directions of the incident and refracted rays respectively. The difference, $\vec{S}_1 - \vec{S}_o$, between these vectors is indicated in Figure 2.3. The length of $\vec{S}_1 - \vec{S}_o$ is the difference between the projections of $\vec{S}_1$ and $\vec{S}_o$ on $\vec{M}_1$. For the case illustrated, $n_1 > n_o$ and therefore $\cos I' > \cos I$. Hence, since Γ is a positive number for Figure 2.3,

$$\Gamma = n_1 \cos I' - n_o \cos I = -n_o \cos I + n_1 \left[\left(\frac{n_o}{n_1} \cos I \right)^2 - \left(\frac{n_o}{n_1} \right)^2 + 1 \right]^{1/2} . \tag{4}$$

Equations (3) and (4) are used in the derivation of the skew ray formulae included in Section 5.

2.3 THE LAW OF REFLECTION

2.3.1 Diagram for reflection. Figure 2.4 shows a ray reflected from a surface. Just as in Figure 2.1, the interface is shown as a straight line, although in general it is a curve. Generally, there will also be a refracted ray which is more or less absorbed as it traverses the medium to the right of the interface. For clarity, only the incident and reflected rays are shown. The calculation of the refracted energy is discussed in Section 21.2.

2.3.2 Statement of the law of reflection. The law of reflection is also stated in two parts:

1. The incident ray, the reflected ray, and the normal to the surface all lie in the same plane.
2. The angle of incidence is numerically equal to the angle of reflection.

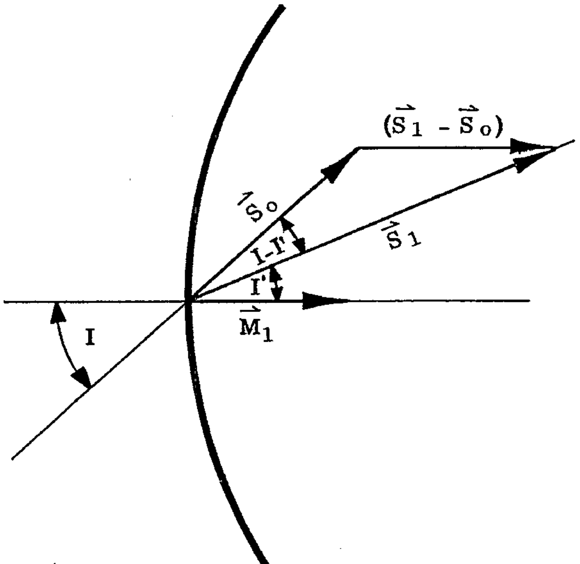


Figure 2.3 – Relation between $\vec{S}_o$, $\vec{S}_1$, and their difference.

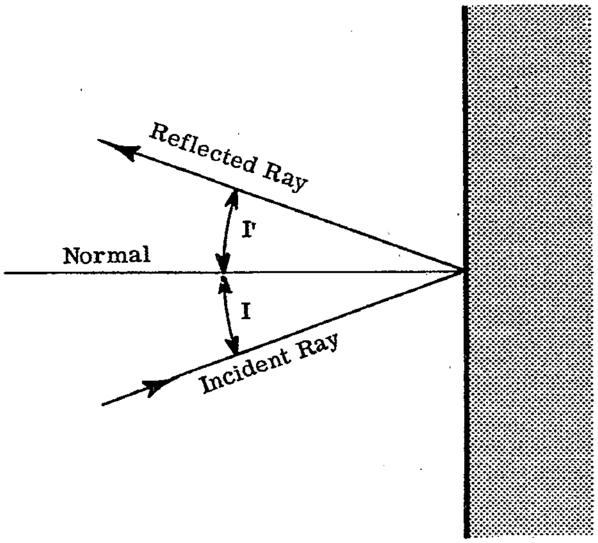


Figure 2.4 - Illustration of reflection.

Note that if I' is labelled as shown in Figure 2.4, then I' is negative while I is positive according to the sign convention. The law of reflection then is

$$I = -I' . \tag{5}$$

2.3.3 Unification of the laws of reflection and refraction. A very convenient way to unify the laws of reflection and refraction is to use the single equation (1) for the law of refraction and to say that in the case of reflection

$$n_1 = -n_o . \tag{6}$$

With this convention, Equation (1) leads directly to Equation (5). This convention will be used later to provide a completely unified treatment of reflection and refraction problems.

2.4 TOTAL INTERNAL REFLECTION

2.4.1 The critical angle. An inspection of Equation (1) shows that if $n_1 < n_o$, and I' is 90°, the angle of incidence then would be given by

$$\sin I_C = \frac{n_1}{n_o} , \tag{7}$$

where I_C is called the critical angle. If the angle of incidence exceeds the critical angle, the reflected ray has associated with it all the incident energy, as though the interface were a perfect mirror. This effect is used to an advantage in the design of prism systems to obtain reflectivity with very little loss of energy. (See Section 13).

2.4.2 Table of critical angles and indices. Table 2.1 lists the critical angle* corresponding to various indices of refraction. These data are useful in the design of prism systems, where it is necessary to be sure that the prism totally reflects all the desired rays.

n	I_c (radians)	n	I_c (radians)	n	I_c (radians)
1.50	0.729728	1.57	0.690526	1.64	0.655753
1.51	0.723820	1.58	0.685308	1.65	0.651099
1.52	0.718020	1.59	0.680177	1.66	0.646517
1.53	0.712324	1.60	0.675132	1.67	0.642005
1.54	0.706730	1.61	0.670168	1.68	0.637562
1.55	0.701234	1.62	0.665286	1.69	0.633186
1.56	0.695834	1.63	0.660481	1.70	0.628875

Table 2.1 - Table of critical angles (n vs I_c).

2.5 INDEX OF REFRACTION

2.5.1 Absolute index of refraction. It is appropriate at this time to discuss the meaning of index of refraction, referred to as n. The absolute refractive index of a material is defined as the ratio of the velocity of light in a vacuum to that in the material,

$$n_o = \frac{v_{vac}}{v_o} . \tag{8}$$

2.5.2 Relative index of refraction. In practice the absolute index of refraction is never directly measured. Instead the velocity in the material is compared to the velocity in air. From this comparison the relative index of refraction can be determined. The relative index of one material with respect to another is equal to the ratio of the absolute indices. For example, the relative index of a substance with respect to air is

$$(n_o)_{rel} = \frac{n_o}{n_{air}} = \frac{v_{vac}/v_o}{v_{vac}/v_{air}} = \frac{v_{air}}{v_o} .$$

---

* As indicated here the angle is expressed in radians. In the future, if an angle is given in radians, the word "radian" will be omitted; if the angle is given in degrees, the degree sign (°) will be used.

Equation (1), which is the basic equation applying to a ray as it traverses a boundary, can be applied without knowing the absolute indices n_0 and n_1. Only the relative index, n_1/n_0, is needed. Hence all refraction problems involve only a ratio of two indices and it is not necessary to know the absolute index of optical materials. Therefore, unless specifically stated, the indices of refraction of optical materials relative to air are used, and it is these relative indices which are measured. (See Section 25.7.3). In problems involving vacuum the absolute index of refraction of air must be used to calculate the absolute index of the material.

2.5.3 Table of refractive indices. The index of refraction of several optical materials is shown in Table 2.2. Except for silicon, where the index applies to the infrared, the indices are for the visible spectrum. Detailed refractive index data on optical glasses are available in catalogs from glass manufacturers. (See paragraph 2.7.9). Materials other than glass are available and are used for optical elements. Refractive index and other data on these materials are discussed in Section 17. It should be noted that the indices given in Table 2.2, as well as in other references, are not only functions of wavelength, which is discussed in Section 2.6, but are also functions of temperature and pressure. The pressure dependence becomes of major importance in the case of gases; sometimes a particular gas at relatively high pressure is used to enclose part or all of an optical system.

Material	n
Vacuum	1.
Air	1.0003
Water	1.33
Fused quartz	1.46
Borosilicate crown glass	1.51
Ordinary crown glass	1.52
Canada balsam	1.53
Light flint	1.57
Dense barium crown	1.62
Extra dense flint	1.72
Silicon (in the infrared)	3.4

Table 2.2 - Refractive indices of various materials.

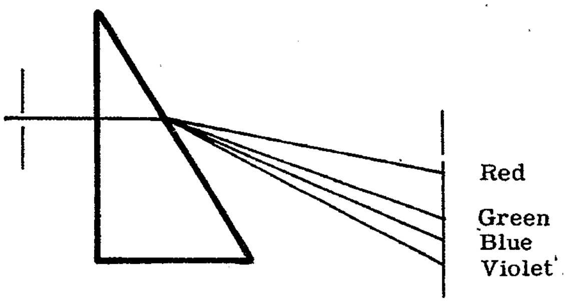


Figure 2.5 - Beam of white light passing through a dispersing prism.

2.6 DISPERSION OF LIGHT

2.6.1 General. It was shown by Newton that white light is not to be considered as a fundamental type, but is rather a composite mixture which can be separated into a range of colors, - that is a spectrum -, by passage through a prism as shown schematically in Figure 2.5. According to the wave theory of light, each color corresponds to a definite frequency of vibration or, when the light is traveling in a vacuum, to a definite wavelength (λ). The shorter waves correspond to the violet end of the spectrum; the longer, to the red. Further investigation has shown that the radiation spectrum extends to longer wavelengths beyond the red, the infrared (IR) region, and to shorter waves beyond the violet, the ultraviolet (UV) region.

2.6.2 Variation of index with wavelength.

2.6.2.1 Since index is inversely proportional to the velocity of light in a given medium, and since this velocity is not constant for all colors, the index is a function of the color of the light. The color may be specified either by stating the frequency or wavelength in vacuum; hence, the index may be considered a function of either frequency or wavelength. Which functional dependence is used depends on the specific problem involved. In geometrical optics, since spectrum lines are used to measure indices, and since these lines are indicated by wavelength (instead of frequency), it is customary to use the functional dependence on wavelength.

2.6.2.2 For a given refracting medium, the absolute refractive index takes on a different value for each wavelength. In all practical cases it is higher for short wavelengths, and lower for long ones. Thus in Figure 2.5 a ray of composite light is incident normally on the first surface. Since the angle of incidence on this surface is zero, the angle of refraction is also zero and the ray is undeviated. At the second surface, however, the light is deviated, the blue ray being bent more than the red. This unequal refraction is called dispersion. The variation of index with wavelength, for most optical materials in the wavelength region where they are used, is such that the index decreases as the wavelength increases. The index varies approximately linearly with $1/\lambda^2$ where λ is the wavelength of the radiation.

2.6.3 Fraunhofer lines. In optical design work, the indices of refraction of the media to be used must be known in the wavelength region in which the device is to be used. (Methods of measuring index will be discussed in Section 24.6). Within the region, the choice of wavelengths at which measurements are made depends partly on convenience in measurement, and partly on custom. The section on glass characteristics applies generally to the visible region. Similar considerations apply to the ultraviolet and infrared regions, but the use of specific wavelengths for reference in those regions is not yet so well established. The range of visible wavelengths runs from about 0.380μ to about 0.740μ. (See Section 4.5). Within this region several reference wavelengths are used which, for historical reasons, are known as Fraunhofer lines, and are customarily denoted by letters assigned to them in a system originated by Joseph von Fraunhofer in his studies of the solar spectrum. In Table 2.3 are given the wavelengths of light of some of the Fraunhofer lines, and the elements from which the lines result. Also included are two additional lines, one in the near infrared, the other in the near ultraviolet, which are being used as standard wavelengths for index measurements.

Color of light	Line	Wavelength, Microns	Element
Infrared		1.0140	Hg
Red	A'	0.7665	K
Red	C	0.6563	H
Yellow	D	0.5893	Na
Yellow	d	0.5876	He
Green	e	0.5461	Hg
Light Blue	F	0.4861	H
Blue	g	0.4358	Hg
Dark Blue	G'	0.4340	H
Violet	h	0.4047	Hg
Ultraviolet		0.3650	Hg

Table 2.3-Fraunhofer and other standard lines.

2.7 CHARACTERISTICS OF OPTICAL GLASS

2.7.1 Reference indices. In designing chromatically corrected systems, it is necessary to make provision for the variation of the index of refraction with wavelength. This will be expanded in later sections, but for now it is important to be aware of the terms and quantities which are usually sufficient to describe the properties of an optical medium in the visible spectrum. In this and in the following paragraphs, reference should be made to specification MIL-G-174, Optical Glass, to become acquainted with approved standard requirements for the military. It is impractical to treat simply the infinite number of indices corresponding to all the wavelengths in white light. Common practice is to select a convenient wavelength near the middle of the eye's sensitive range, using one which can be easily and accurately reproduced. The refractive index of the material at this wavelength is then used as a basic reference both in design and in material designation. The material's refractive index for yellow light corresponding to the mean wavelength of the two sodium D lines is usually used in the United States and is designated n_D . European practice is to use n_d , the index corresponding to the yellow helium line. Similarly, the terms n_F and n_C are the indices of refraction for the F and C lines of hydrogen and provide reference indices in the blue and red regions.

2.7.2 Abbe constant. A commonly used expression for identifying chromatic properties is the Abbe constant, which is defined as

$$\nu = \frac{n_D - 1}{n_F - n_C} .$$

The symbol V , rather than the Greek ν is frequently used; however ν will be used in this text. The Abbe constant is named for its inventor, the German scientist Ernst Abbe. It is often called the nu value or the vee number. The numerator, $n_D - 1$, is called the refractivity for the sodium D lines.

2.7.3 Partial dispersion. The difference between any two indices for a given substance, corresponding to two different wavelengths, is called the partial dispersion. Hence $n_D - n_C$ is the partial dispersion for the D and C lines. The particular partial dispersion, $n_F - n_C$, is called the mean dispersion because it covers approximately the visual range of wavelengths. Use is sometimes made of a partial dispersion ratio, for example, $(n_D - n_C)/(n_F - n_C)$.

2.7.4 Glass type number. It has become common practice to identify a glass by the type number, which is a six-digit number. The first three digits of the type number are the first three rounded digits of the refractivity, $(N_D - 1)$, and the last three digits of the type number are the first three rounded digits of the ν- value of the glass. A glass with $n_D = 1.51250$ and $\nu = 60.5$ would have a type number of 513605.

2.7.5 Staining. In addition to the quantities involving refractive indices, which have been mentioned above, additional optical characteristics must be considered in optical design. One of these, surface staining, obviously affects the transmittance; such staining is accelerated by the presence of acidic atmospheres, for example caused by carbon dioxide or perspiration. Staining can be measured quantitatively by the time required to form a film one quarter of a wavelength thick when the sample is immersed in nitric acid under controlled conditions of concentration and temperature.

2.7.6 Dimming. A characteristic somewhat related to staining is surface dimming, which occurs when the polished sample is exposed to moist air. It can be measured quantitatively by exposing the sample to a 100% relative humidity atmosphere at a given temperature for a specified time, and classifying the appearance of the surface.

2.7.7 Bubbles. All glasses contain some bubbles, or inclusions, varying in size and number according to the glass type. A glass sample is classified according to the number of bubbles in a specified volume of material. If a bubble is less than 0.02 mm in diameter (or some other standard value), it is not counted as it is considered invisible.

2.7.8 Table of optical glass characteristics. Table 2.4 lists the quantities described above in identifying glass. The glass type number is given in both the extreme left and extreme right hand columns. The second column at the left gives the ν - number. There follow eleven columns giving the refractive index for the corresponding wavelengths. The next column gives the mean dispersion. There follow six columns listing two numbers for each glass type. The one in large type is a partial dispersion, the other a partial dispersion ratio. The specific gravity is listed in the next column; as the metal parts of optical instruments become more and more fabricated of light alloys, the glass weight becomes an important factor and must be considered in overall optical design. The next column gives the staining time in hours, and adjacent to it is listed the stain test class. In the next column is given the dimming test class number, running from 1 (not visibly dimmed) to 5 (dimming interfering with clear vision). The bubble code is given in the next to the last column; the code runs from 1 (few bubbles) to 4 (many bubbles). The letter P following a glass type indicates that this type is available in a form which makes it resistant to gamma rays and X-rays. The term fine annealed indicates that permanent strain on cooling has been virtually eliminated.

2.7.9 Availability of glass tables. Designers, or interested students should obtain from glass manufacturers the latest catalog information. Some suggested sources are: (1) in the United States, Bausch and Lomb, Rochester, New York; Corning Glass Works, Corning, New York; Eastman Kodak Co., Rochester, New York; Hayward Glass Co., Whittier, California; Pittsburgh Plate Glass Co., Pittsburgh, Pennsylvania; and (2) abroad, Chance-Pilkington Optical Works, St. Asaph, England; Tozai Boeki Kaisha, Ltd., No. 13, 4-Chome, Shiba-Tamuracho, Minatoku, Tokyo, Japan; Minex, P.W.O. Works, Jelenia Góra, Poland; Ohara Optical Glass Manufacturing Co., Sagamihara, Kanagawa, Japan; Parra-Mantois, Le Vésinet, France; Schott Glass Works, Mainz, West Germany; Schott Glass Works, Jena, East Germany. Catalogs of Russian manufacturers are published by Gosudarstvennoe Isdatelstvo, Moscow, USSR. Additional U.S. companies and representatives of foreign companies are listed in the Optical Industry Directory (See page 1-5).

CHARACTERISTICS — OPTICAL GLASS

Indices Given are for "Fine Annealed" Glass

TYPE	V $\frac{n_D-1}{n_F-n_C}$	1014.0	$n_{A'}$ Potassium 766.5	n_C Hydrogen 656.3	n_D Sodium 589.3	n_d Helium 587.6	n_e Mercury 546.1	n_F Hydrogen 486.1	n_g Mercury 435.8	$n_{G'}$ Hydrogen 434.1	n_h Mercury 404.7	365.0	n_F-n_C	$n_C-n_{A'}$ $\frac{n_C-n_{A'}}{n_F-n_C}$	n_D-n_C $\frac{n_D-n_C}{n_F-n_C}$	n_e-n_D $\frac{n_e-n_D}{n_F-n_C}$	n_F-n_e $\frac{n_F-n_e}{n_F-n_C}$	$n_{G'}-n_F$ $\frac{n_{G'}-n_F}{n_F-n_C}$	$n_h-n_{G'}$ $\frac{n_h-n_{G'}}{n_F-n_C}$	Specific Gravity	Staining Time at 25°C—Hrs	Stain Test Class	Dimming Test Class	Bubble Code	TYPE
BOROSILICATE CROWN													**BOROSILICATE CROWN**												
498670	**67.0**	*	1.49316	1.49577	**1.49800**	1.49808	1.49984	1.50320	1.50717	1.50732	1.51048	*	0.00744	0.00261 0.351	0.00223 0.300	0.00184 0.247	0.00336 0.452	0.00412 0.554	0.00316 0.425	2.44	+100	1	2.5	1	498670
506596 (See Note 1)	**59.6**	*	1.50058	1.50347	**1.50600**	1.50609	1.50811	1.51196	1.51656	1.51675	1.52039	*	0.00849	0.00289 0.340	0.00253 0.298	0.00211 0.248	0.00385 0.454	0.00479 0.564	0.00364 0.429	2.47	+100	1	2.5	2	506596
511635	**63.5**	1.5017	1.50578	1.50860	**1.51100**	1.51107	1.51300	1.51665	1.52096	1.52114	1.52454	1.53057	0.00805	0.00282 0.350	0.00240 0.298	0.00200 0.248	0.00365 0.453	0.00449 0.558	0.00340 0.422	2.48	+100	1	1.0	1	511635
517645 517645P	**64.5**	1.5076	1.51179	1.51461	**1.51700**	1.51707	1.51899	1.52262	1.52690	1.52709	1.53043	1.53644	0.00801	0.00282 0.352	0.00239 0.298	0.00199 0.248	0.00363 0.453	0.00447 0.558	0.00334 0.417	2.53	+100	1	1.0	1	517645 517645P
CROWN													**CROWN**												
513605	**60.5**	1.5030	1.50708	1.50999	**1.51250**	1.51258	1.51459	1.51846	1.52304	1.52324	1.52685	1.53332	0.00847	0.00291 0.344	0.00251 0.296	0.00209 0.247	0.00387 0.457	0.00478 0.564	0.00361 0.426	2.51	+100	1	1.5	1	513605
518596	**59.6**	1.5083	1.51242	1.51544	**1.51800**	1.51807	1.52015	1.52413	1.52886	1.52906	1.53279	1.53951	0.00869	0.00302 0.348	0.00256 0.295	0.00215 0.247	0.00398 0.458	0.00493 0.567	0.00373 0.429	2.54	+100	1	3.0	3	518596
523586	**58.6**	1.5130	1.51729	1.52036	**1.52300**	1.52307	1.52520	1.52929	1.53415	1.53435	1.53819	1.54505	0.00893	0.00307 0.344	0.00264 0.296	0.00220 0.246	0.00409 0.458	0.00506 0.567	0.00384 0.430	2.53	+100	1	3.0	4	523586
524595	**59.5**	*	1.51838	1.52140	**1.52400**	1.52408	1.52618	1.53021	1.53500	1.53520	1.53898	*	0.00881	0.00302 0.343	0.00260 0.295	0.00218 0.248	0.00403 0.457	0.00499 0.567	0.00378 0.428	2.53	+100	1	3.0	1	524595
LIGHT BARIUM CROWN													**LIGHT BARIUM CROWN**												
541599 541599P	**59.9**	1.5309	1.53522	1.53833	**1.54100**	1.54109	1.54323	1.54736	1.55226	1.55246	1.55633	1.56326	0.00903	0.00312 0.346	0.00267 0.296	0.00223 0.247	0.00413 0.457	0.00512 0.567	0.00387 0.429	2.84	+100	1	2.5	1	541599 541599P
573568	**56.8**	*	1.56614	1.56954	**1.57250**	1.57259	1.57498	1.57962	1.58514	1.58538	1.58975	*	0.01008	0.00340 0.337	0.00296 0.294	0.00248 0.246	0.00464 0.460	0.00576 0.571	0.00437 0.434	3.20	22	2	3.0	3	573568
573574 573574P	**57.4**	1.5616	1.56619	1.56956	**1.57250**	1.57259	1.57497	1.57953	1.58497	1.58521	1.58951	1.59723	0.00997	0.00337 0.338	0.00294 0.295	0.00247 0.248	0.00456 0.457	0.00568 0.570	0.00430 0.431	3.21	9	3	2.0	3	573574 573574P
DENSE BARIUM CROWN													**DENSE BARIUM CROWN**												
588612	**61.2**	*	1.58184	1.58513	**1.58800**	1.58811	1.59036	1.59474	1.59992	1.60015	1.60424	*	0.00961	0.00329 0.342	0.00287 0.298	0.00236 0.245	0.00438 0.455	0.00541 0.562	0.00409 0.425	3.29	.03	5	4.0	1	588612
611572	**57.2**	1.5993	1.60423	1.60785	**1.61100**	1.61109	1.61364	1.61853	1.62438	1.62462	1.62923	1.63754	0.01068	0.00362 0.339	0.00315 0.295	0.00264 0.247	0.00489 0.458	0.00609 0.570	0.00461 0.432	3.57	0.033	5	2.0	4	611572
611588	**58.8**	*	1.60439	1.60793	**1.61100**	1.61109	1.61357	1.61832	1.62396	1.62421	1.62867	*	0.01039	0.00354 0.341	0.00307 0.295	0.00257 0.247	0.00475 0.457	0.00589 0.567	0.00446 0.429	3.58	0.017	5	3.0	4	611588
612595	**59.5**	*	1.60544	1.60896	**1.61200**	1.61209	1.61455	1.61924	1.62484	1.62508	1.62946	*	0.01028	0.00352 0.342	0.00304 0.296	0.00255 0.248	0.00469 0.456	0.00584 0.568	0.00438 0.426	3.45	.02	5	4.0	1	612595
617549	**54.9**	1.6048	1.60995	1.61370	**1.61700**	1.61710	1.61977	1.62493	1.63115	1.63142	1.63634	1.64516	0.01123	0.00375 0.334	0.00330 0.294	0.00277 0.247	0.00516 0.459	0.00649 0.578	0.00492 0.438	3.66	0.050	5	1.5	4	617549
617551	**55.1**	*	1.50994	1.61371	**1.61700**	1.61710	1.61976	1.62490	1.63104	1.63130	1.63617	*	0.01119	0.00377 0.337	0.00329 0.294	0.00276 0.247	0.00514 0.459	0.00640 0.572	0.00487 0.435	3.36	0.2	4	3.5	3	617551
620603	**60.3**	*	1.61342	1.61696	**1.62000**	1.62011	1.62255	1.62724	1.63282	1.63306	1.63748	*	0.01028	0.00354 0.344	0.00304 0.296	0.00255 0.248	0.00469 0.456	0.00582 0.566	0.00442 0.430	3.58	**	5	4.5	3	620603
623569	**56.9**	*	1.61606	1.61978	**1.62300**	1.62309	1.62571	1.63073	1.63675	1.63701	1.64171	*	0.01095	0.00372 0.340	0.00322 0.294	0.00271 0.247	0.00502 0.458	0.00628 0.573	0.00470 0.429	3.58	**	5	4.0	1	623569
638555	**55.5**	*	1.63074	1.63461	**1.63800**	1.63810	1.64084	1.64611	1.65243	1.65271	1.65772	*	0.01150	0.00387 0.337	0.00339 0.295	0.00284 0.247	0.00527 0.458	0.00660 0.574	0.00501 0.436	3.59	**	5	4.0	3	638555
651558	**55.8**	*	1.64362	1.64757	**1.65100**	1.65109	1.65389	1.65924	1.66563	1.66591	1.67097	*	0.01167	0.00395 0.338	0.00343 0.294	0.00289 0.248	0.00535 0.458	0.00667 0.572	0.00506 0.434	3.81	**	5	4.0	1	651558

Note 1: Available in Condenser quality only.
*Data not currently available.

** Dissolves in HNO_3

Courtesy of **BAUSCH & LOMB OPTICAL CO.**

Table 2.4 - Excerpt from commercial glass catalog.

3 CONSIDERATIONS OF PHYSICAL OPTICS

3.1 INTRODUCTION

3.1.1 Diffraction nature of optical images.

3.1.1.1 The goal in designing a lens system on the basis of geometrical optics is to find a combination of lenses for which all rays in a specified cone of rays that diverges from an object point P are converged upon the corresponding image point P' such that the optical paths of all rays from P to P' are equal. Other requirements are added. For example, it may be required that points P and P' shall belong to a single object plane and a single image plane, respectively. Even when the design satisfies all these requirements to a high degree, the image P' of a self-luminous object point P is not a point but consists of a central bright spot surrounded by systematically distributed dark and bright fringes whose contour and width depend upon the contour and dimensions of the aperture of the lens. If, for example, the lens aperture is circular and if the self-luminous object point is located upon or near the optic axis, the image consists of a circular, central bright spot surrounded alternately by dark and bright rings. The central bright spot is called the Airy disk. Its diameter decreases as the diameter of the lens aperture is increased. The actual image of the object point is modified to such a degree by diffraction from the finite lens aperture that this image is appropriately called a diffraction image.

3.1.1.2 The diffractive nature of the image may not be so apparent with, for example, high-speed objectives in which compromises among the geometrical corrections and tolerable aberrations must be made. However, the image will generally exhibit effects due to diffraction, i. e., effects that cannot be explained from Snell's law of refraction or reflection alone. In any case, the image of a point will not be a point; an exact point by point similarity between object and image cannot be achieved. Resolution of details in the image of the object is restricted first by the degree of correction of the optical system and finally by the laws of diffraction, i. e., by the laws governing the bending of light rays from the paths consistent with Snell's law of refraction and reflection.

3.1.1.3 Whereas the action of most optical systems can be explained by the principles of geometrical optics, the action of other systems such as phase microscopy can be understood only as a proposition in diffraction. However, in any system, the ultimate resolving power and contrast in the fine-grained details of an image are determined by diffraction.

3.1.2 Diffraction and interference.

3.1.2.1 Broadly, diffraction is the phenomenon whereby waves are modified in direction, amplitude, and in phase by interaction with an object or obstacle. In its most general sense, diffraction includes the phenomena of refraction and reflection but these two phenomena are ordinarily considered apart from diffraction. However, when the dimensions of the object become comparable to the wavelength, the concepts of refraction and reflection become useless. With such small objects, even scattering becomes a direct aspect of diffraction.

3.1.2.2 Interference is the process by which two or more overlapping waves interact so as to re-enforce one another in some regions and to oppose one another in other regions. This process is essentially one of addition of the instantaneous amplitudes of the overlapping waves. It matters a great deal whether or not the overlapping waves are coherent. In case the added waves are incoherent, the time-averaged energy density is simply the sum of the time-average of the energy density associated with each wave, i. e., the resulting energy follows the law of superposition of energy. Conversely, it may be concluded that if the time-average of the energy densities follows the law of superposition of energy, the interfering waves are essentially incoherent. Interference includes the process by which a given wave is split or decomposed into two or more waves (often called component waves). These component waves are automatically coherent since they belong to the same wave-train. The action of interferometers can usually (but not always) be explained adequately by considering the sum of two or more waves.

3.1.2.3 Diffraction and interference are related processes, but diffraction is the more inclusive. In fact, diffraction effects can include interference effects as special cases. For example, in explaining the "interference fringes" produced with monochromatic light leaving two small pinholes that are illuminated coherently from a third pinhole, it is natural to regard the formation of the interference fringes as an interference effect, i. e., as a process of adding the two well defined spherical waves that emerge from the pair of pinholes. However, as the area of the pinholes is increased, the location of the origin of the spherical waves that leave different portions of the pinholes begins to matter. The process of summing the effects of the infinite many wavelets that leave the pinholes is now carried out most conveniently by means of integrals that characterize diffraction processes.

3.2 THE PHYSICAL NATURE OF LIGHT

3.2.1 The wave theory

3.2.1.1 Much evidence supports the view that light is propagated as electromagnetic waves whose wavelengths λ fall in the visible range from 0.38 to 0.76 microns. The transverse nature of electromagnetic waves is illustrated in Figure 3.1 in which E and H denote the electric and magnetic vectors, respectively. The electric

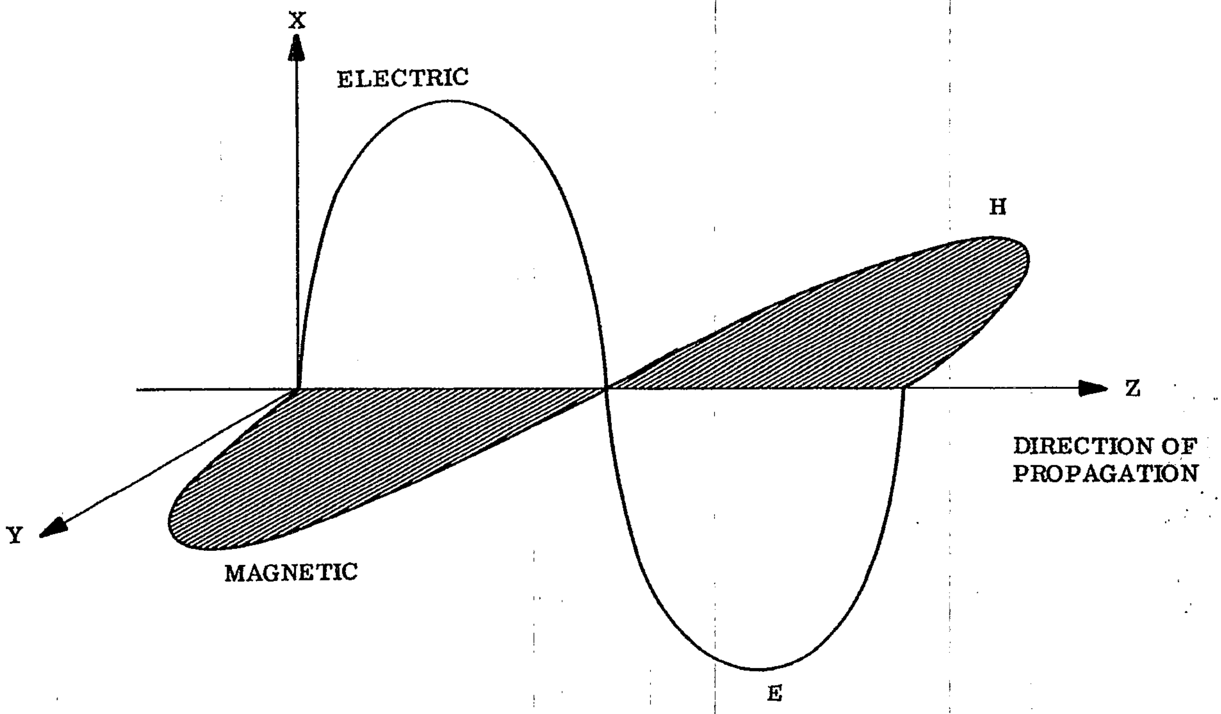


Figure 3. 1—The electromagnetic nature of a plane polarized light wave. The electric vector E and the magnetic vector H oscillate at right angles to the direction of propagation and at right angles to one another.

and magnetic vectors are ordinarily perpendicular to each other and to the direction of propagation. The electric vector describes an electric force field that will cause an electric charge to vibrate along the E-direction. Thus, the electric vector produces displacements of ions or electrons along the positive or negative E-direction, respectively. The vectors E and H are inseparable and are mutually dependent. For this reason it usually suffices to specify only the electric vector. The luminous flux can be computed whenever the radiant flux of the electromagnetic waves is known (as it is when the E-vector is specified).

3.2.1.2 The velocity of all electromagnetic waves in vacuum is a constant = c = 299792.5 kilometers per second. The velocity of monochromatic waves in non-vacuum media invariably depends upon the wavelength and is accordingly called the phase velocity to distinguish it from the group velocity of a group of monochromatic waves. The refractive index n of a medium is defined such that

$$n = \frac{\text{velocity in vacuum}}{\text{phase velocity in the medium}} \qquad (1)$$

Let T denote the period of vibration of a monochromatic wave. Let $\nu = 1/T$ denote the frequency ν of vibration. Then if v denotes the phase velocity

$$v = \nu \lambda = \frac{c}{n} \qquad (2)$$

As an electromagnetic wave moves from one medium into another, its frequency remains fixed. Hence its wavelength must change such that the wavelength λ in a medium of refractive index n varies according to the law

$$\lambda = \frac{c}{n\nu} = \frac{cT}{n} = \frac{\lambda_o}{n} \qquad (3)$$

where $\lambda_o = cT$ = wavelength in vacuum.

3.2.2 Plane-polarized light waves.

3.2.2.1 A plane-polarized light wave is one whose electric vector vibrates in a fixed plane (which we shall call the plane of polarization) in homogeneous media that do not rotate the plane of polarization. The wave illustrated in Figure 3. 1 is plane-polarized. If the direction of propagation is the Z-axis, the magnitude E (z, t) of the electric vector can be specified as the trigonometric function

$$E\ (z,\ t) = a \cos\ (knz + \phi - \omega t) \tag{4}$$

where

z = distance measured along Z
t = time
k = $2\pi/\lambda$
ω = $2\pi/T$
λ = wavelength
T = period for one complete vibration

ϕ = phase angle
n = refractive index. It can be a function of z for variable media.
a = amplitude of the wave. It is an exponential decreasing function of z for absorbing media.

The phase angle ϕ is needed for specifying the phase of one wave relative to another. If, for example,

$$E_1 = a_1 \cos\ (knz + \phi_1 - \omega t) \tag{5}$$

$$E_2 = a_2 \cos\ (knz + \phi_2 - \omega t) \tag{6}$$

the corresponding waves differ in phase by the amount $\phi_1 - \phi_2$ at like values of t and z.

3.2.2.2 The state of vibration or polarization is the same for all points that belong to a wavefront. On each wavefront

$$knz + \phi - \omega t = \text{constant} = w \tag{7}$$

where w is different for each wavefront. The wavefront moves so as to satisfy Equation (7). By differentiating the members of Equation (7) with respect to the time t, one finds that

$$\frac{dz}{dt} = v = \frac{\omega}{kn} = \frac{1}{n}\frac{\lambda}{T} = \frac{c}{n^2};\ \frac{\lambda}{T} = \frac{c}{n} \therefore \frac{1}{n} \cdot \frac{c}{n} = \frac{c}{n^2} \tag{8}$$

3.2.2.3 The wavefronts of the plane-polarized wave described by Equation (4) are perpendicular to the Z-axis, the direction of propagation. If the plane-polarized plane wave is propagated along an arbitrary direction OP, Figure 3.2, the magnitude E of the electric vector assumes the form

$$E = a \cos \left[kn\ (px + qy + rz) + \phi - \omega t \right] \tag{9}$$

where p, q and r are the direction cosines of OP with respect to X, Y, and Z, respectively. Thus,

$$p^2 + q^2 + r^2 = 1. \tag{10}$$

Equation (9) reduces to Equation (4) when the direction of propagation OP is the Z-direction only, for then $p = q = o$ and $r = 1$. It is important to observe that the wave motion of Equations (4) and (9) is of the form

$$E = a \cos\ (\Phi - \omega t) \tag{11}$$

where

$$\Phi = kn\ (px + qy + rz) + \phi \tag{12}$$

with p, q, and r defined as the direction cosines of the direction of propagation of the plane-polarized, plane wave. The electric vector vibrates in the wavefront.

3.2.3 Energy in a single wave. The instantaneous energy, W_i (whether energy flux or energy density) in the wave is proportional to E^2, where E denotes the instantaneous magnitude of the electric vector. We take the factor of proportionality as unity and write from Equation (11)

$$W_i = E^2 = a^2 \cos^2 (\Phi - \omega t). \tag{13}$$

The oscillations of light waves are so rapid that the eye or other known detectors are unable to follow the in-

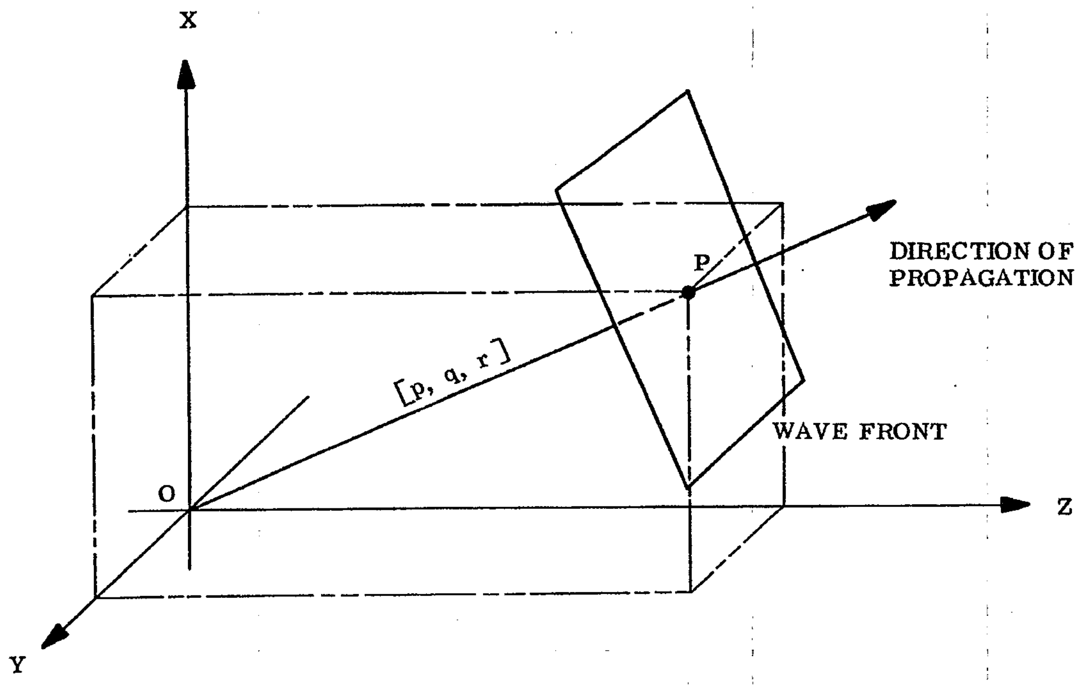


Figure 3. 2—Notation with respect to the propagation of a plane wave.

stantaneous values. Rather, the time average W of W_i is detected and measured. It suffices to average over one period T of oscillation. Thus,

$$W = \frac{1}{T}\int_o^T a^2 \cos^2(\Phi - \omega t)\,dt$$

$$= \frac{a^2}{2T}\int_o^T \left[1 + \cos 2(\Phi - \omega t)\right] dt. \tag{14}$$

Since $\omega = 2\pi/T$, it follows almost directly that

$$\int_o^T \cos 2(\Phi - \omega t)\,dt = 0. \tag{15}$$

Hence,

$$W = a^2/2, \tag{16}$$

i.e. the time-averaged energy density or energy flux in a single wave is proportional to the square of its amplitude. W is independent of, for example, the phase angle ϕ of the single plane wave.

3.3 INTERFERENCE BETWEEN WAVES

3.3.1 Collinear, coherent waves.

3.3.1.1 Two waves will be called collinear when they are propagated in the same direction. We consider the interference of two plane-polarized,* plane waves that are propagated in the same direction with a constant phase difference δ . The magnitudes E_1 and E_2 of the electric vectors of two unlike plane waves assume from Equation (11) the form

---

* The electric vectors of these two plane polarized waves are assumed parallel, i.e., are assumed to vibrate in the same fixed plane.

$$E_1 = a_1 \cos (\Phi_1 - \omega t) \; ; \qquad E_2 = a_2 (\Phi_2 - \omega t) \; . \tag{17}$$

From Equation (12)

$$\Phi_1 - \Phi_2 = \delta \; , \tag{18}$$

the phase difference between the two waves.

3.3.1.2 Let E denote the magnitude of the electric vector formed by the sum of E_1 and E_2, i.e., formed by the interference of the two waves. Then,

$$E = a_1 \cos (\Phi_1 - \omega t) + a_2 \cos (\Phi_2 - \omega t) \; . \tag{19}$$

Let W be the time-averaged energy density formed by the two interfering waves. As in paragraph 3.2.3,

$$\begin{aligned} W &= \frac{1}{T} \int_o^T E^2 \, dt \; . \\ &= \frac{a_1^2}{T} \int_o^T \cos^2 (\Phi_1 - \omega t) \, dt + \frac{a_2^2}{T} \int_o^T \cos^2 (\Phi_2 - \omega t) \, dt \\ &\quad + \frac{2 \, a_1 a_2}{T} \int_o^T \cos (\Phi_1 - \omega t) \cos (\Phi_2 - \omega t) \, dt \\ &= \frac{a_1^2}{2} + \frac{a_2^2}{2} + 2 \, a_1 a_2 \, I \end{aligned} \tag{20}$$

where

$$I = \frac{1}{T} \int_o^T \cos (\Phi_1 - \omega t) \cos (\Phi_2 - \omega t) \, dt \; . \tag{21}$$

But

$$\cos (\Phi_1 - \omega t) \cos (\Phi_2 - \omega t) = \frac{1}{2} \left[\cos (\Phi_1 + \Phi_2 - 2\omega t) + \cos (\phi_1 - \phi_2) \right] . \tag{22}$$

As in Equation (15),

$$\int_o^T \cos (\Phi_1 + \Phi_2 - 2\omega t) \, dt = 0 \; .$$

Hence,

$$I = \frac{\cos (\Phi_1 - \Phi_2)}{2T} \int_o^T dt = \frac{\cos (\Phi_1 - \Phi_2)}{2} = \frac{\cos \delta}{2} \tag{23}$$

Finally, from Equations (23) and (20) we find that the time-averaged density, W , produced by the interference of two, plane-polarized, collinear, plane waves having amplitudes a_1 and a_2 and phase difference ($\phi_1 - \phi_2$) is

$$W = \frac{1}{2} \left[a_1^2 + 2 \, a_1 a_2 \cos \delta + a_2^2 \right] . \tag{24}$$

3.3.1.3 For constructive interference, the phase difference $\Phi_1 - \Phi_2 = \delta$ between the two waves is 0, 2π , 4π , etc., so that

$$W = \frac{1}{2} (a_1 + a_2)^2 \; . \tag{25}$$

For destructive interference, $\delta = m\pi$ where m is an odd integer. Correspondingly,

$$W = \frac{1}{2} (a_1 - a_2)^2 \; . \tag{26}$$

It should be noted from Equation (26) that $W = 0$ when the two waves have equal amplitudes and are out of phase. Thus, two plane waves that are propagated in the same direction can cancel one another everywhere, or they can re-enforce one another everywhere provided that their phase difference δ is a suitably chosen constant. The time-averaged energy density of the resultant wave is not merely the sum of the time-averaged energy densities of the two separate waves except in the special cases $\cos\delta = 0$. See Equations (25) and (16). The waves are coherent when δ is constant.

3.3.2 Collinear, incoherent waves.

3.3.2.1 One should expect that when light or any other radiation from two independent sources overlap, the resulting energy density is simply the sum of the overlapping energy densities, i.e., the law of superposition of energy should apply. The interfering waves ought to be incoherent. The following somewhat oversimplified argument brings to bear the essential physics underlying the interference of incoherent waves.

3.3.2.2 The time-averaged energy density, produced by two interfering waves that have amplitudes a_1 and a_2 and the phase difference δ, is given by Equation (24). We shall avoid considering the sum of a large number of waves having randomly distributed phase differences δ (as will occur with independent sources) by supposing that in a short interval of time the phase differences δ between the two interfering waves are distributed with equal probability in the interval $0 \leqq \delta \leqq 2\pi$. Then from Equation (24)

$$W = \frac{1}{2}\left[a_1^2 + 2\,a_1 a_2\,\overline{\cos\delta} + a_2^2\right] \tag{27}$$

where $\overline{\cos\delta}$ is the average value of $\cos\delta$ when all values of δ are equally probable in the interval $0 \leqq \delta \leqq 2\pi$. One can show that

$$\overline{\cos\delta} = 0 \ . \tag{28}$$

In this manner we conclude that

$$W = \frac{1}{2}\left(a_1^2 + a_2^2\right) \tag{29}$$

so that the interference between incoherent waves is of that degenerate variety to which the law of superposition of energy applies.

3.3.3. Non-collinear, coherent waves.

3.3.3.1 The theory of paragraph 3.3 is almost but not quite adequate for explaining and interpreting the interference fringes that appear in Twyman Green and other double-beam interferometers; for in these interferometers the mirrors are usually tilted so that the two interfering waves are not propagated in the same direction. It is well known that a series of straight and parallel interference fringes are seen when the interfering waves are not collinear and when the reflecting surfaces are optical flats.

3.3.3.2 We may suppose without essential loss of generality that one wave is propagated along the direction OP that makes any angle θ with Z but is oriented so that the direction cosine $q = 0$. The two interfering waves are described by Equation (17); but $\Phi_1 - \Phi_2$ will not be given by Equation (18). Instead,

$$\begin{aligned}\Phi_1 &= knz + \phi_1 \\ \Phi_2 &= kn\,(x\sin\theta + z\cos\theta) + \phi_2\end{aligned} \tag{30}$$

so that

$$\Phi_1 - \Phi_2 = \phi_1 - \phi_2 - knx\sin\theta + knz\,(1 - \cos\theta) \ . \tag{31}$$

From Equations (20) and (23) the time-averaged energy density formed by the two interfering, coherent waves is

$$W = \frac{1}{2}\left[a_1^2 + 2\,a_1 a_2 \cos(\Phi_1 - \Phi_2) + a_2^2\right] . \tag{32}$$

Substituting $\Phi_1 - \Phi_2$ from Equation (31) and setting $\phi_1 - \phi_2 = \delta$, the fixed phase difference between the two waves, one obtains

$$2W = a_1^2 + a_2^2 + 2\,a_1 a_2 \cos\left[\delta - knx\sin\theta + knz\,(1 - \cos\theta)\right] \tag{33}$$

in which θ is the angle indicated in Figure 3.3, $k = 2\pi/\lambda$ and n is the refractive index of the medium. δ is the phase difference between the two interfering waves having amplitude a_1 and a_2 at the point $x = 0$, $z = 0$.

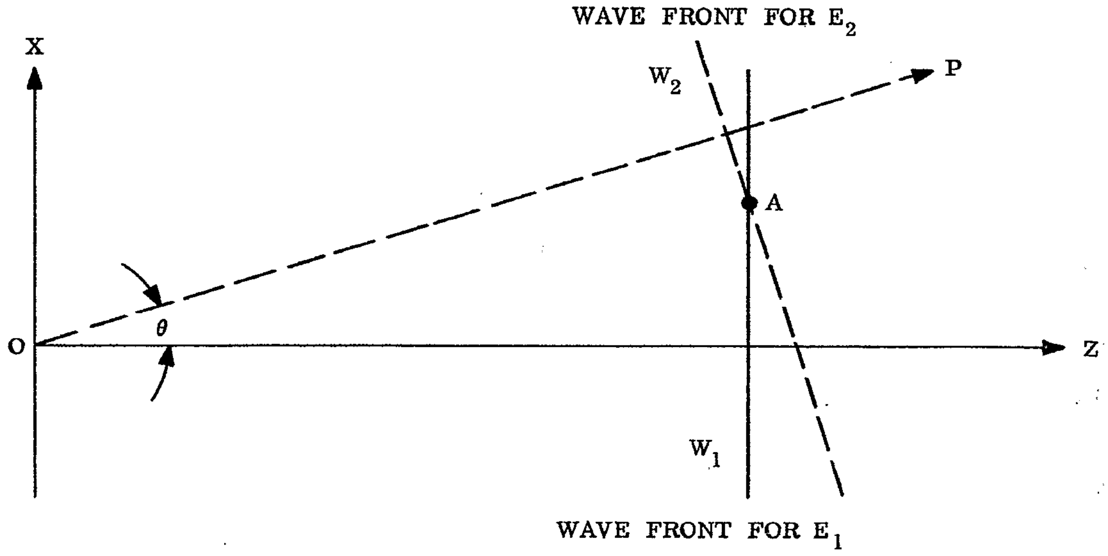


Figure 3.3— Interference between two plane wavefronts W_1 and W_2 that are propagated along different directions.

3.3.3.3 In double beam interferometry, the angle θ is usually so small that one can set $\sin\theta = \theta$ and $1 - \cos\theta = \theta^2/2$. If, then, one makes observations in planes z near $z = 0$, Figure 3.3, the term containing z in Equation (30) can be neglected. The approximation thus obtained is the usual interference formula

$$2W = a_1^2 + a_2^2 + 2a_1 a_2 \cos(\delta - 2\pi nx\theta/\lambda) . \tag{34}$$

The fringes are repeated whenever x is increased by an amount Δx such that

$$kn\Delta x \sin\theta = 2\pi .$$

The fringe width h is therefore given by

$$h = \Delta x = \frac{\lambda}{n \sin\theta} . \tag{35}$$

The greater fringe widths belong to the longer wavelengths.

3.3.3.4 In case the fringes are photographed with a camera that images a plane into a plane, the interference fringes will be straight. Suppose, however, that the camera has curvature of field. In this case a plane z = constant will not be focused upon the photographic plate. Consequently, one has to expect from Equation (33) that the photographed fringes will be curved and that the distortion of the fringes should increase as θ and z are increased.

4 VISUAL OPTICS

4.1 INTRODUCTION

4.1.1 Characteristics of the human eye. The design of an efficient optical instrument must include consideration for the use of the instrument. When the human eye is to be the translating instrument, the instrument must be designed for proper seeing. This section will call attention to some of the advantages and limitations of the human eye and seeing that are important for instrument design. The human eye is sensitive to radiant energy from 380 to about 740 $m\mu$ in wavelength. The limits of visibility for young eyes are about 313 - 900 $m\mu$, but for practical purposes the narrower range is adequate and representative for average eyes. Light is defined as radiant energy evaluated according to its capacity to produce visual sensation. A few quanta can stimulate the retina and be seen as light. To see an object, light of suitable quality (color) and intensity from the object must form an image on the retina of adequate size, contrast, and duration for the retina to transform the light energy into nerve energy, and the nerve impulses must be conducted to the brain and integrated into consciousness. Age, glare, state of adaptation and visual acuity will modify vision.

4.1.2 Seeing. Seeing is a learned ability and training can improve the individuals seeing to limits set by the eye and nervous system. Seeing is a perceptual process that is affected by and incorporates other sensations, emotions, association mechanisms simultaneously active with vision, education, and past experience. It varies with the condition of the individual and the entities must be statistical probabilities of seeing rather than absolute values.

4.1.3 Loss of vision. The eye and vision are disturbed by many conditions and diseases. Emmetropia refers to an average normal eye, ametropia indicates a defective eye and amblyopia an eye with little or no vision that appears normal. Additional defects of the eye are covered in paragraph 4.3.3.

4.2 ANATOMY OF THE EYE

4.2.1 Physcial structure. The human eye, as illustrated in Figure 4.1, is a nearly spherical organ held in shape by a tough, outer, whitish-sclerotic coat and the pressure of its viscous content. The cornea, the transparent front part of the sclera, protrudes slightly as it has a greater curvature. Inside the sclera is the choroid containing the blood vessels, the opaque pigment and the ciliary process. The ciliary process includes the iris and the muscles which focus the lens of the eye. The pupil is the opening in the center of the iris. The retina covers the inside of the choroid to the ora serrata near the ciliary process. The space between the cornea and the iris is called the anterior chamber and between the iris and the lens is a posterior chamber. Both are filled with the aqueous humor. The space back of the lens and ciliary process is filled with the vitreous humor. The lens is attached to the ciliary muscle by many fibers or suspensory ligaments. Except for the opening in the iris the pigmentation of the sclera and iris normally makes the eye light tight. A lack of eye pigmentation is called albinism and vision is impaired by glare from light leakage onto the retina.

4.2.2 Intraocular pressure. The internal pressure of the eye is maintained quite constant by a balance of the formation of the aqueous humor at the back part of the ciliary process, from which it passes out through the pupil into the anterior chamber, and drains through the canal of Schlemm.

4.2.3 Metabolism. The transparent media, cornea, lens and vitreous do not have blood vessels and receive their nourishment from the fluids surrounding them. The transparency of the cornea depends on its relative hydration. The front part of the retina contains blood vessels which furnish nourishment to it and to the adjacent vitreous.

4.2.3.1 The retina is one region of the body where it is possible to see (with the ophthalmoscope) the condition of the blood vascular system and recognize changes from many systemic diseases. The focussing ability of the eye is altered by a change in the blood sugar concentration from inadequately controlled diabetes. Glaucoma is a disease characterized by an increase in the pressure within the eye ball and unless arrested promptly will lead to mechanical damage and loss of sight.

4.2.4 Development. The eye is developed early and is fairly well formed by six weeks after conception. An outgrowth from the front of the brain becomes the optic nerve and the retina of the eye. When this cup-shaped formation nearly reaches the skin of the embryo, that part of the skin sinks below the surface and becomes modified to form the lens of the eye. The skin closes over to form the cornea and the sclera. The choroid and the ciliary process form between the sclera and the retina. Like the brain, the eye is relatively large at birth al-

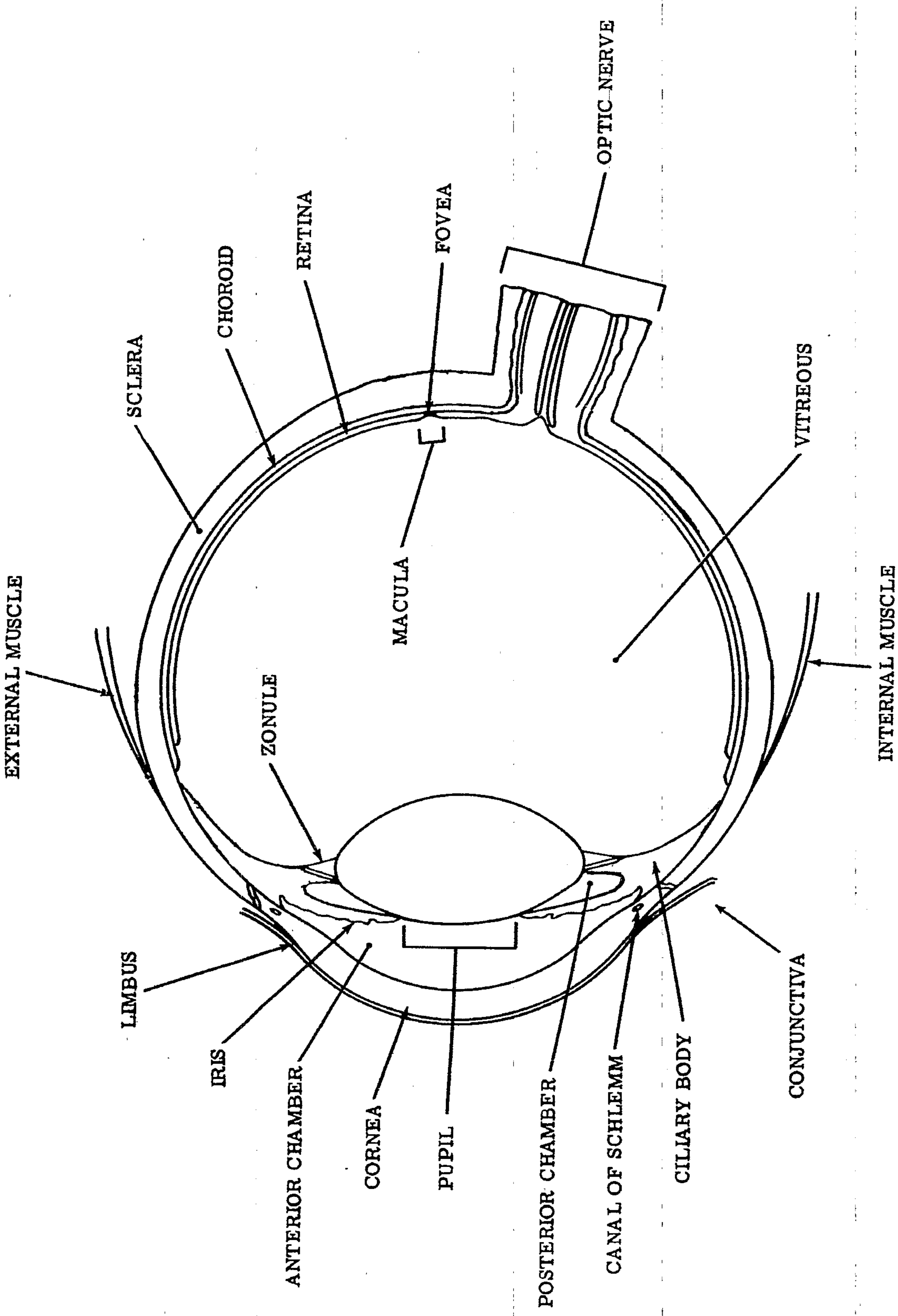


Figure 4.1. - Horizontal section of the right eye.

though vision then is imperfect and improves for several years. Color vision may not reach its greatest sensitivity until in the late teens. The various parts of the eye do not grow at the same rate and the eye and the body do not grow at the same rate. It is remarkable that the regulatory mechanisms tend to balance these different rates of growth to produce the emmetropic eye.

4.2.4.1 The muscles which control the eye will be described later. Briefly however, muscular action is usually a balance between opposing pairs of muscles which contain many contractile units and the resulting movement usually shows the action of the units in a stepwise progression, and fine oscillations when in equilibrium.

4.3 OPTICAL CONSTANTS OF THE EYE

4.3.1 Use of the "standard eye." As one would expect there are no universal dimensions for an eye; one finds instead, considerable variation in all dimensions. A good image formed on the retina may be the result of each part of the eye being perfect in form and refractive index, or the shapes and indices of the parts may have compensated for each others defects. Complete testing of each observer's eye would be time consuming and require special equipment. Instead a "standard" or typical eye is established and used as a standard observer for computational problems. Individual eyes can be examined to discover whether or not they correspond to the standard. There are several systems for "reduced" eyes, and a commonly used set of optical and mechanical characteristics for a typical eye is illustrated in Figure 4.2. Reduced is used here in the sense of an optically equivalent system.

4.3.2 Aberrations. Like other optical systems the eye is subject to the usual aberrations. The coordination of the focussing system and the retinal structure with sunlight over many years evolution has minimized some of the problems. Distortion and field curvature rarely bother in ordinary seeing, and chromatic aberration does not disturb vision. With the small pupils,of 3-4mm and average daylight, spherical aberration is minimal, although in dim light with large pupils it lessens vision.

4.3.3 Corrective lenses. The chief defects of the eye are myopia, hyperopia or hypermetropia, astigmatism, presbyopia and aniseikonia. The hyperopic eye focuses the image of a distant object behind the retina,and the myopic eye in front of the retina. In old age the focussing ability of the lens declines and this condition is termed presbyopia. Astigmatism results from asymmetry of the cornea. Aniseikonia will be discussed in paragraph 4.7 and aphakia will be discussed in paragraph 4.4.

4.3.3.1 Far sightedness, or hyperopia, can be due to the axial length of the eye being too short, or the focussing mechanism too weak, and is corrected by placing in front of the eye a plus lens of proper strength to replace the image on the retina. In near-sightedness, or myopia, the image is formed in the vitreous because the eye is too long, or the focussing mechanism is too strong, and the defect is corrected with a minus spectacle lens. Astigmatism due to irregular curvature of the cornea is corrected by a cylindrical spectacle lens.

4.3.3.2 Spectacles are usually fitted so that the back surface (vertex) of the lens is about 14 millimeters in front of the cornea although minus lenses for myopia may be set closer at 9 to 11 millimeters. Changing the position alters the effective power of the lens. Eyeglasses may be tilted slightly downward 4° to 12° for reading.

4.3.3.3 People with astigmatic corrections must wear their glasses for comfortable vision over long periods when using optical instruments. In recent years optical designers have made oculars with the eye point far enough from the lens so that the individual can see the whole field while wearing spectacles. The distance from the front of the spectacle lens to the cornea can vary from around 17 millimeters to 11 millimeters. If a substitute lens is mounted on the optical instrument to take the place of a spectacle lens, its power must be changed from that of the prescription when the substitute lens will be at a different position from the cornea than the spectacle lens. A substitute lens with cylindrical power must be mounted in proper orientation to the axis of the cylinder so it cannot rotate from the correct position.

4.3.3.4 People with only near or far sightedness (no astigmatism) usually remove their glasses when using optical instruments and refocus the instrument to correct for their defect. Therefore, focussing eyepieces should have sufficient range for the people intended to use them. A range of ± 1 diopter will include about 70 percent; ± 2 diopters will include about 85 percent; and ± 4 diopters about 98 percent of spectacle prescriptions.

4.3.3.5 Critical seeing can take place only when the image is located on the fovea at the center of the macula of the retina, as illustrated in Figures 4.1 and 4.2. This establishes a visual axis which is some 5-7° from the optical axis of the eye. The retina is blind over the area of the optic disc where the nerve fibers enter the eye to distribute over the retina, and this blind spot subtends some 7° vertically and 5° horizontally.

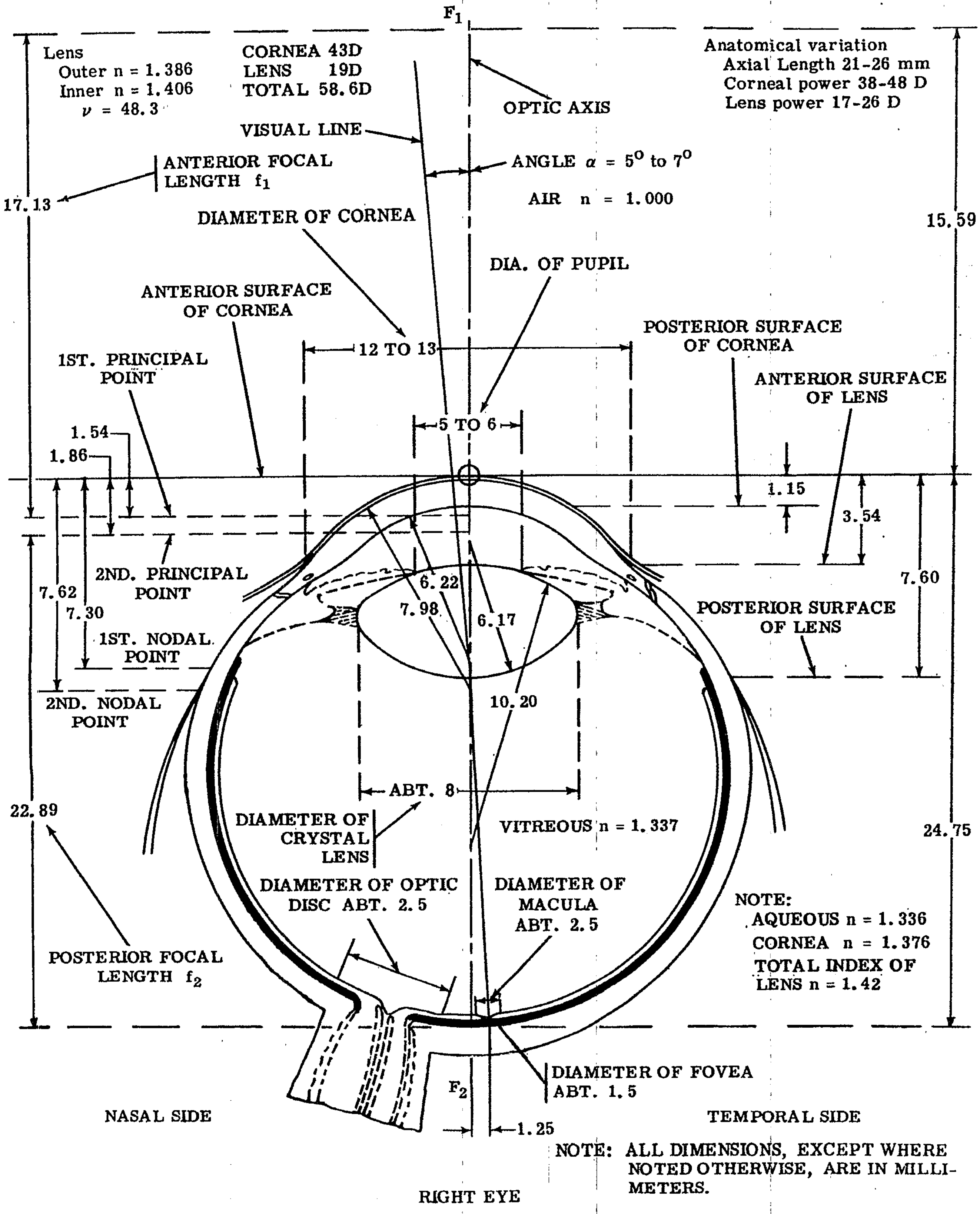


Figure 4.2 - Optical constants for a "standard eye."

4.4 IMAGE FORMATION AND THE RETINA

4.4.1 Cornea. The cornea is the first refracting surface for light entering the eye and is responsible for about 43 of a total of 58 diopters power of the eye. Normally the cornea is transparent and the refracting power is due to the curvature and refractive index difference between it and air on one side, and the aqueous humor on the other. The cornea in size averages 12 millimeters horizontally and 11 millimeters vertically.

4.4.1.1 A change in the hydration of the cornea can affect the light passing through it either by distortion or decreased transparency. The decrease caused by fluids and some early contact lenses, or from changes in old age, scatters light and haloes appear around light sources or small bright objects. Haloes from age changes are rarely reversible.

4.4.1.2 The two surfaces of the cornea usually are of similar curvature and have no lens effect on the entering light. Any deformity of the curvature of the cornea (astigmatism) distorts the image. Such changes are measured with a keratometer (ophthalmometer) and corrected by adding a corresponding cylinder of opposite sign into the spectacle lens for the eye. An extreme elongation of the center of the cornea (keratoconus) can be corrected by contact lenses. Astigmatism has some relation to the tension of the eye muscles and may change slowly from a vertical meridian to a horizontal meridian of greatest curvature during later life. There may be some residual astigmatism as well as that from the corneal surface.

4.4.1.3 Vision specialists sometimes refer to astigmatism with the rule (stronger power vertical) and against the rule (meridian of greatest curvature horizontal) based on the direction of movement of light reflected from the eye during skiascopic refraction.

4.4.1.4 Haidinger's Brushes are seen on looking at the blue sky (polarized), or at a uniform source of polarized blue light, as a diffuse cross. Some observers believe this phenomenon is due to the birefringence of the cornea. Other observers hold that it is due to neural structure or pigment arrangement in the retina. Attempts to use the Brushes for differential diagnosis of eye conditions has been unsuccessful so far.

4.4.2 Pupil. The pupil is the opening in the center of the iris as illustrated in Figures 4.1 and 4.2. In dim illumination the pupil opens to about 8 millimeters diameter in young eyes, and closes to about 2 millimeters diameter in intensely bright light. Under average conditions the pupil has a diameter of 3.5 to 4 millimeters. Resolution of the eye is decreased when the pupils are much larger or smaller than 3 to 4 millimeters. With ageing, the pupil remains smaller, and in extreme old age may not be more than 2 to 3 millimeters. The pupil is a stop, or diaphragm, in the dioptric system of the eye that affects image formation, illumination of the retina and the aberrations of the system. With small pupils (2 millimeters or less) diffraction becomes important.

4.4.2.1 The iris is composed of radial and circular muscle fibers and the size of the pupil is a resultant of these antagonistic muscles. Consequently the pupil shows continuous fine fluctuations in size, as well as opening and closing with changed luminance. The iris is not under voluntary control. Convergence of the eyes to a closer point in space also closes the pupil and this increases the depth of field.

4.4.2.2 Stimulation of the cornea, conjunctiva or eyelids, causes a slight dilation, followed by contraction of the pupil. Strong sensory stimulation, fear, and pain cause dilation via the psycho-sensory reflex. Many drugs effect the size of the pupil and some are used in the medical treatment of the eye to dilate (mydriasis) or contract (myosis) the pupil. Normally, both pupils respond together from the stimulation of either eye although the sizes may not be exactly the same. A marked difference in sizes indicates disease.

4.4.2.3 The pupil can decrease from 8 to 3 millimeters in 4 to 5 seconds. Dilation of the pupil from 3 to 6 millimeters takes 5 to 10 seconds and maximum dilation may take 5 to 10 minutes. Contraction at 5.5 to 7 millimeters per second and dilation at 3.0 to 4.5 millimeters per second is reported.

4.4.2.4 In designing optical instruments for visual use it should be kept in mind that the usable part of the exit pupil is no larger than the pupil of the eye. In order to decrease the precision with which the eye must be placed at the exit pupil in viewing, it is sometimes advisable to design the instrument so that the diameter of the exit pupil is considerably larger than any possible diameter of the pupil of the eye. In this case the portion of the exit pupil transmitting light to the observer's retina is limited to the size of the eye pupil, and the usable diameter of the entrance pupil (for axial bundles of rays) is equal to the diameter of the eye pupil multiplied by the magnification. However, if the exit pupil is smaller than the pupil of the eye the light entering the eye is limited by the exit pupil, and in instruments requiring maximum illumination on the retina every attempt should be made to provide an exit pupil diameter as large as the largest possible diameter of the pupil of the eye. Average pupil size for age and luminance are shown in Figure 4.3.

4.4.3 Lens. The lens of the eye changes curvature to focus light onto the retina. The lens is a transparent elastic body with an outer capsule, a less dense cortex, and a denser inside core. The lens is held in position

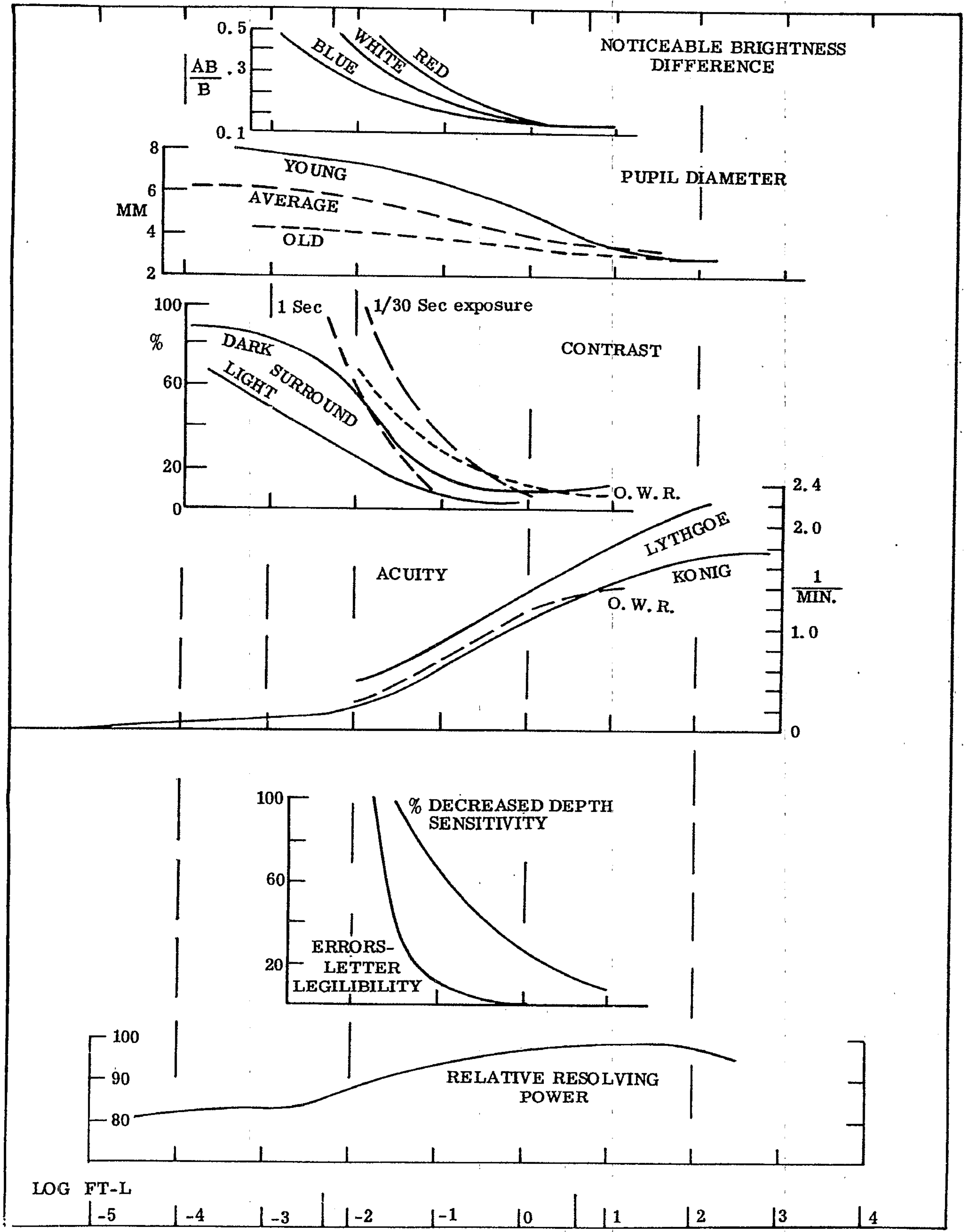


Figure 4.3. - Variation of some attributes of vision with Luminance.

by the suspensory ligaments, as shown in Figure 4.1. The ciliary process has circular, radial, and oblique muscle fibers which contract to pull on the fibers of the zonule and flatten the lens; or relax to lessen the tension and let the lens bulge to a more spherical form. Continuous fluctuations from muscle action take place producing amplitudes of 0.1 diopters focal change with a frequency of 4 to 8 cycles per second and smaller frequencies of 2 and 0.3 cycles per second. The lens has a total refracting power of some 19 diopters and the amplitude of accommodation of the lens varies from some 15 diopters in children to about 0.5 diopter in old age. The depth of field is about 0.5 diopter. However, to focus the eye from near to far requires 0.7 to 0.8 second, far to near 0.4 to 0.5 second, and near to far and back to near 1.15 to 1.25 seconds. When vision is less than 20/20, when exophoria exceeds 8 prism diopters at 33 centimeters, or when myopia, hyperphoria, or astigmatism are present, the time required to focus the eye will increase from that mentioned above.

4.4.4 Accommodation. The curvature of the front and back surfaces of the lens are different and the front surface is said to be hyperbolic in young people. The focussing of the lens is controlled by the sympathetic nervous system and cannot be altered voluntarily. There have been two theories advanced regarding accommodation. Helmholz thought that relaxation of the tension on the suspensory ligaments permitted the elastic lens substance, which had been deformed in the unaccommodated state, to return to its more convex form. E. F. Fincham's experiments indicate that such relaxation allows a highly elastic capsule to deform the lens substance from its unaccommodated state to the greater convexity required. The variations in thickness in different parts of the capsule favors the latter theory.

4.4.4.1 When the eye sees only an empty field lacking detail the lens tends to focus, not at the 20 foot "infinity" of the vision specialists, but at about 1 meter. This near-sightedness is called empty field myopia for a bright empty field, and night myopia when the empty field is due to darkness. In the latter case the change in spherical aberration from the dilated pupil and the Purkinje shift also contribute to the total myopia. Changes in the curvature of the lens can be measured objectively from changes in the Purkinje-Samson images reflected from the surfaces of the lens, or with an optometer from changes in the retinal image, using either light or invisible infrared radiation.

4.4.4.2 At about forty years of age the focussing mechanism begins to gradually fail (presbyopia) and additional plus lens correction becomes necessary to see details at the usual reading distance. The lens also tends to become yellowish, blues are seen less well in old age, and less light gets to the retina. In some eyes the lens becomes opaque (cataract) and must be removed to restore vision. The eye lacking a lens is said to be aphakic and the spectacle lens correction must be increased to substitute for the lens. As the spectacle lens has a fixed focus the aphakic eye will be corrected only at one distance. When one eye is aphakic and the other is not, the difference in the size of the images on the retina precludes binocular vision.

4.4.4.3 Optical instruments with focusable eyepieces must be designed to have an adequate adjustment in power to permit older people to use them, and to provide at least -2 diopters when designed for night use.

4.4.4.4 The vitreous humor is a transparent gel of slightly greater refractive index than water, that fills the space between the lens and ciliary process and the retina. Sometimes particles of tissue (muscae volitantes) tend to hang or float in the vitreous and are seen when one is observing through optical instruments. These may be fragments left over as the vitreous formed, or that have broken away during life. Nothing can be done to remove these fragments and they should be ignored. In some diseases, parts of the vitreous become opaque and vision is lost to a corresponding extent.

4.4.4.5 The retina, covering most of the area behind the ciliary process, translates light energy into nervous energy and contains the first coordinating nerve cells in the visual system. The front part facing the lens is composed of blood vessels, nerve cells and fibers and connective tissues, and at the back of the retina are the light sensitive rod and cone cells and protective pigment layer. The entrance of the optic nerve forms a disc (a blind spot where there are no light sensitive cells) and the visual angles subtended by this disc are about 7° and 5° as illustrated in Figure 4.2. The disc is about 3.5 millimeters (15.5° to center) on the nasal side of the optical pole of the eye and 1.5 degrees below the horizontal meridian of the eye.

4.4.4.6 The retina thins at the visual axis, some 5° temporal to the optical pole, as there are no blood vessels or nerve fibers over the fovea. The macula subtends about 12° and is 2.5 to 3 millimeters in diameter. The fovea includes about 1.5 millimeters of the center of the macula, or about 5° of subtended arc and is the most sensitive part of the retina. Some anatomists recognize an area of about 0.35 millimeters in the center of the fovea called the foveola.

4.4.4.7 The center of the fovea only contains cones and those at the central region are longer, thinner and more densely packed than cones elsewhere in the retina. This rod-free area is about 0.5 millimeter in diameter and subtends about 50 minutes of arc. From here to the edge of the retina the number of cones per unit area decreases, and the number of rods increases. At 20°, as illustrated in Figure 4.4, the rod population is densest.

4.4.4.8 The sensitivity of the retina to light varies with the area stimulated as shown in Figure 4.5. The fovea is most sensitive and used for seeing fine detail and color. Color sensitivity varies with position on the retina.

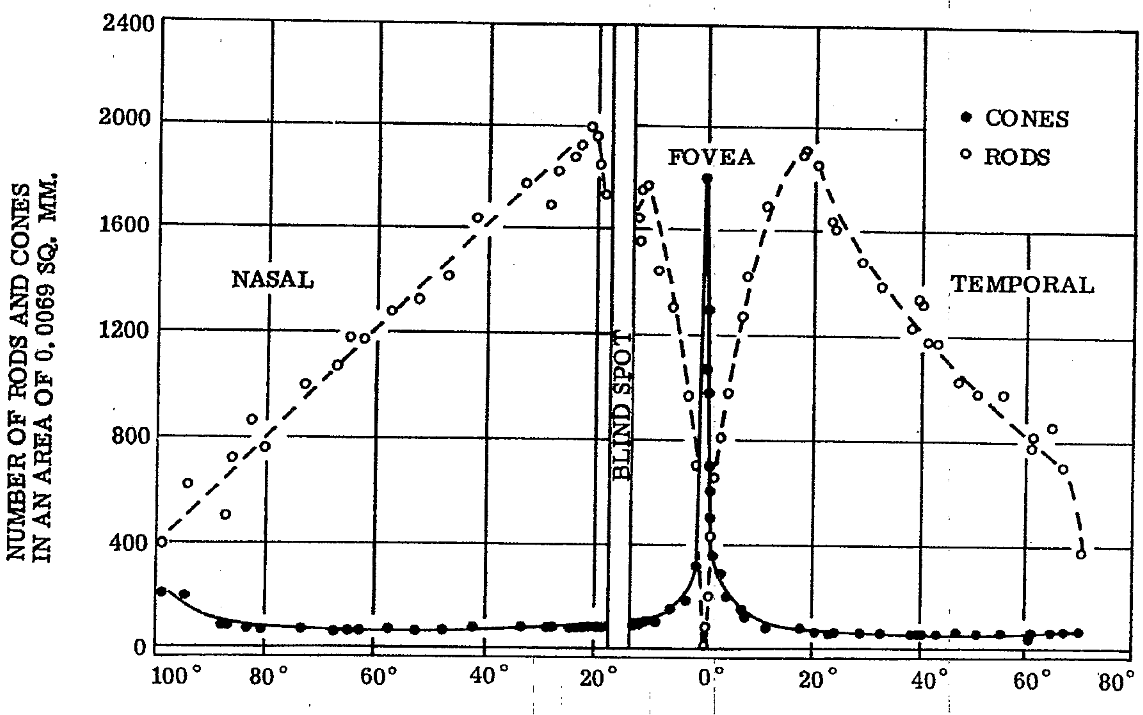


Figure 4.4 - The distribution of cones and rods across the retina (horizontal meridian).
(From National Research Council, A. Chapanis 1949)

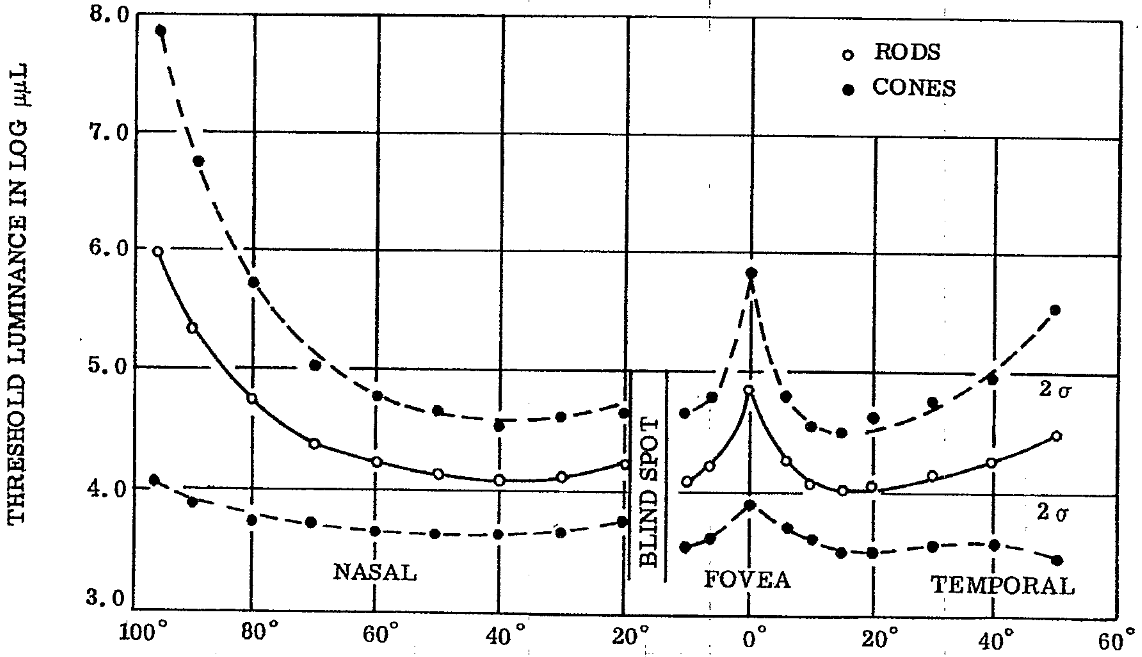


Figure 4.5 - Sensitivity to just perceptable luminance across the retina
(From National Research Council, A. Chapanis 1949)

4.4.4.9 The rods contain rhodopsin, which is bleached by light, and the products formed stimulate nerve conduction. Rods are sensitive to very small amounts of light and operate from a few quanta to a luminance of about that of moonlight (0.01 ft-L). The cones contain iodopsin and have a useful range from about 0.006 ft-L to 10,000 ft-L. Vision with the rod cells at low levels of light is called scotopic and cone cell vision at high levels is called photopic. The overlapping region (0.1-0.01 ft-L) is called mesoptic vision. The structure of the rods and cones is complex and the exact mechanism of vision is not fully known. A nerve fibre conducts or it does not. Nerve fibers respond to stimulation after a latent period and are insensitive during the refractive period following conduction. Chemical action and electrical potentials accompany the impulse. These factors and the light intensity establish the timing of the impulses. The frequency rate of conduction, and the interconnections of the nerve cells, codes the light from the image on the retina into the brain and consciousness. The cones of the fovea are individually connected to a single nerve fiber and have a direct path into the optic nerve. Beyond the fovea, the rods and some cones are connected in groups by the retinal nerve cells, thereby facilitating pattern vision.

4.4.4.10 The nerve fibers from the right half of each eye cross at the point where the optic nerves join, and go to the right hemispheres of the brain. Those from the left halves of each retina go to the left hemisphere. What is seen in the right half of each visual field is connected to the left hemisphere of the cerebrum and vice versa. Cutting one optic nerve would blind that eye while damage to an optic tract would blind the same half of both eyes.

4.4.5 Resolution. The rods and cones give the retina a mosaic structure that determines resolution. Minimum resolution depends on three factors: retinal location of the image as illustrated in Figure 4.6; the nature of the image and the criterion used; and adequate time for stimulation. A very small light (bright on dark) will be seen when its image has enough quanta (2-8) to stimulate the retina, and the smallness of the bright spot depends solely on its brightness. Two small dark objects can be recognized as two when their images spread over or involve two cones providing the diffraction patterns are sufficiently separated. The arc subtense of a cone is about 1 minute (49 to 73 seconds from a gradient of 4 to 6μ for the cones) and the average eye resolves details subtending 1 minute of arc at the eye (70μ at 250 millimeters). An extended image (rather than point) can be seen when much smaller. For example, a telephone wire can be seen against the sky when it subtends only 0.5 second. Horizontally or vertically oriented wires are seen about equally well, but when at 60° or 120° to the horizontal they are only about one-third as visible. A break in a line, or the misalignment of two lines, one above the other, (e.g. scale and vernier) of 4 seconds is visible. Grating objects have different

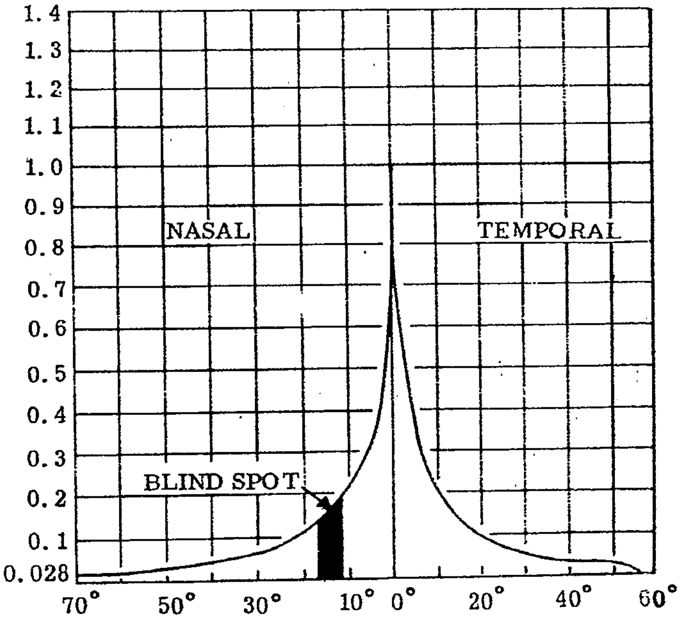


Figure 4.6 - Distribution of visual acuity across the retina expressed in degrees from the fovea

(From National Research Council, A. Chapanis 1949) thresholds. The minimum separable for a grating in motion is reported to be about 2 minutes for a visual acuity of 1.0 for a 2° retinal area and an optical nystagmus criterion. Resolving power decreases with distance from the fovea, to 25% at 5° and only 7% of foveal resolution at 10° from the fovea. Thresholds, as illustrated in Figure 4.7, decrease linearly as the distance from the fovea increases.

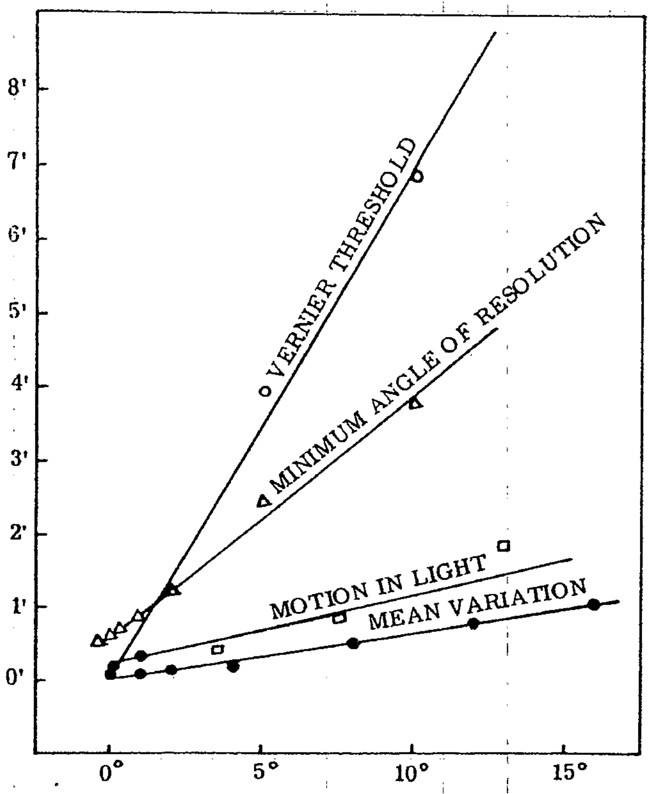


Figure 4.7 - Threshold decrease with distance (in degrees) from the fovea.
(From American Journal of Optometry No. 46, F.W. Weymouth, 1958)

4.4.5.1 Light entering the center of the pupil is more effective than light entering the edge of the pupil. This Stiles-Crawford effect is explained by the orientation of the cones within the retina since the effect occurs only in photopic vision (see paragraph 4.4.4.9). At about 1 millimeter from the center of the pupil there is a decrease to about 90%, at 2 millimeters 70%, 3 millimeters 40% and at 4 millimeters from the center of the pupil the effectiveness of the light is about 20% of that passing through the center of the pupil.

4.4.5.2 The light on the retina varies with the area of the pupil. The Troland (formerly called photon) is the unit of intensity of stimulus for 1mm^2 of pupil area and a luminance of 1c/m^2. Luminance (mL) times 5d ^{2}/2 = Trolands, when d is the pupil diameter in millimeters. Correction may be required for the Stiles-Crawford effect and for the transparency of the eye should a value other than 0.5 be preferred.

4.4.5.3 Optical instruments for visual use should be designed to provide the best image on the retina, of a size and intensity resolvable by the retina. When measurements or judgments can be made by vernier acuity they will be most sensitive, e.g. when a scale value can be aligned to the specimen, the measurement will be more accurate than if the scale is superimposed on the specimen. Small linear detail is more readily seen when imaged horizontally or vertically on the retina, rather than at oblique angles.

4.5 SEEING

4.5.1 Sensitivity. Light of equal energy from different parts of the spectrum does not appear equally bright to the eye as illustrated in Figure 4.8. The yellow-green at 555mμ is brightest and is ten times brighter than the blue of 470 or the red of 650mμ. The standard observer curve represents an internationally accepted sensitivity for use in calculations involving color and relative sensibility of the eye. Like the reduced eye discussed in paragraph 4.3.1, it is representative of average eyes and exact agreement is rarely found between it and an individual eye. Sensitivity curves for individual eyes reveal small departures from the standard observer curve that were averaged out of the standard.

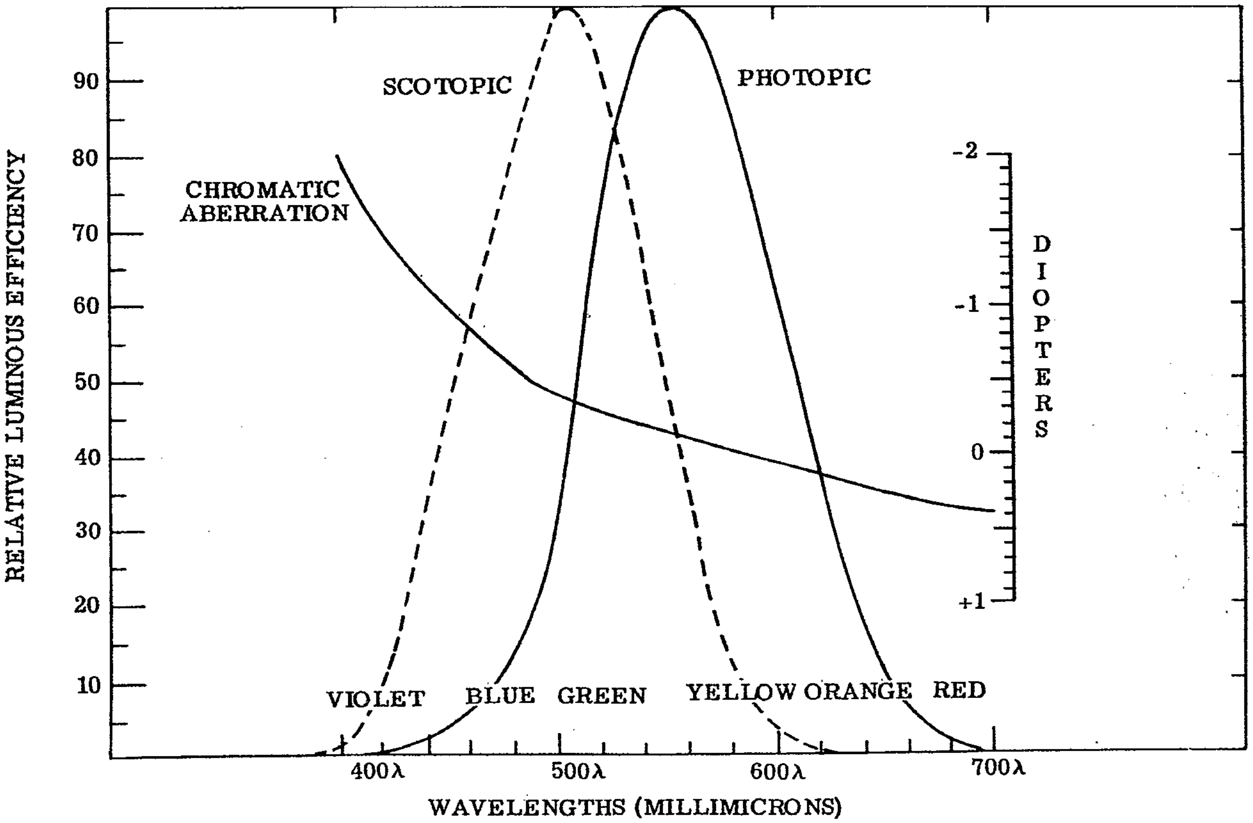


Figure 4.8. - Photopic and Scotopic standard observer curves and chromatic aberration of the eye.

4.5.2 Contrast and time. The eye can adapt itself to see over a wide range of light. The changes within the eye which make this possible involve the pigments of rods and cones and probably neural factors. The sensitivity of an eye in darkness increases rapidly for a few minutes, followed by a gradual increase for about ten minutes as illustrated in Figure 4.9. A further rapid increase of sensitivity (decrease of threshold) takes place until equilibrium is reached. While a further slow increase in sensitivity may take place for hours the amount is not large after one hour in the dark. The curve of Figure 4.9 is typical, and the change after ten minutes marks the end of the cone adaptation and the beginning of the dark adaptation of the rods. The shape of the curve depends on the adaptation state of the retina at the beginning of the dark period. The eye should be exposed for some minutes to a known light (12 log $\mu\mu$L) before measurement. This adaptation may be measured as the threshold at a given time, or as the time required to reach a known sensitivity. Wearing red glasses ($\lambda > 590$mμ) accomplishes some adaptation without being in total darkness.

4.5.2.1 After adaptation, the eye is more sensitive to blueish-green at 510mμ, and the scotopic standard observer curve applies as illustrated in Figure 4.8. The change in the brightest region of the spectrum, from 555 to 510mμ, is called the Purkinje shift. In the mesoptic range, as the eye becomes dark adapted, blues appear brighter and reds darker until color vision fails at about 0.04ft-C of illumination.

4.5.2.2 Dark adaptation is effected by the amount of previous exposure and the physical condition of the individual. It is facilitated to a limited extent by an increase in the available oxygen and is decreased by malnutrition (especially vitamin A deficiency), some drugs, and various diseases. Night-blind individuals cannot adapt to lower light intensities and are disqualified from night operations. When the luminance is too low for the sensitivity of the cones, one has to look to one side of an object so that its image is not on the fovea. The retina is more sensitive for scotopic vision at about 20^{o} from the fovea. This coincides with the greatest density of the rods.

4.5.3 Flicker. When the eye is illuminated by brief flashes of light, alternated with darkness, the eye sees a flickering until the rate reaches 10 to 30 cycles per second when the images fuse and appear continuous. This rate of fusion is called the critical flicker frequency (CFF) and slightly different values are obtained from increasing the rate than from decreasing the rate to fusion. The CFF increases with increased luminance. Talbot's law states that, "fluctuating and steady lights of the same energy content appear equally bright," although recent experimentation indicates that for brief exposures intermittent light is less efficient, while for long exposures fluctuations help. The difference is probably related to the small fluctuating movements of the eye. A great many factors affect the CFF and attempts to use it as a criterion of vision or health have not been very satisfactory.

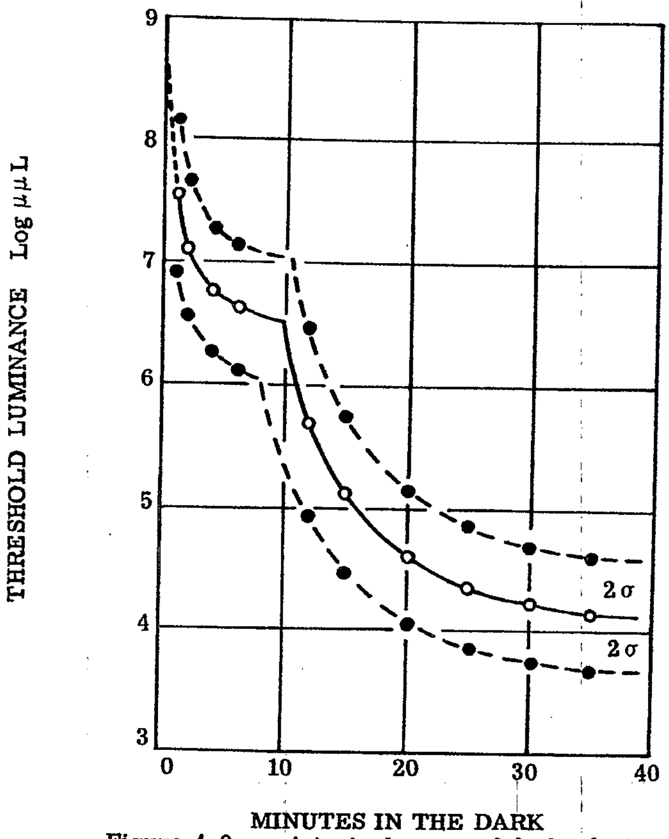


Figure 4.9. - A typical curve of dark adaption.

(From National Research Council, A. Chapanis 1949)

4.5.3.1 When the image is stabilized on the same exact part of the retina, vision gradually fades and disappears. Continuous small fluctuating movements (30-80 cps) and slow drifting of the eye prevents loss of vision. After the image drifts too far from its original position, a quick motion returns the image to the more sensitive part of the retina. To avoid the effect of eye movements in vision research, it is necessary that the stimulus be exposed no longer than 1/100th of a second. During steady fixation for 3 to 4 seconds the image may move over 25 to 50 receptors.

4.5.4 Measuring vision. For the practical purposes of measuring vision for the prescription of spectacle lenses various types of test charts are used, usually consisting of letters of different sizes. The standard is a 5 minute square letter, the individual details of the letter subtending at the observer's eye 1 minute of arc. The reference line on the chart is made with details of a size for the viewing distance to be used. Ordinary Snellen letter charts are designed for use at 20 feet from the observer. Other lines on the chart have graded sizes of letters, e.g. the line marked 40 ft. on the chart would subtend details of 2 minutes at the eye. Visual acuity (VA) is expressed as a fraction, the numerator of which is the design distance for the chart (usually 20 ft.) and the denominator is the line which can be read at that distance. With such a chart 20/20 vision would be normal, 20/15 would be better than normal, and 20/80 would be about 1/4 normal vision (observers only able to read at 20 feet, the line normal observers would read at 80 feet). These charts have high contrast black on a white background. In Europe similar charts are based on 6 meters distance (very nearly 20 feet) and the corresponding acuities are written as 6/6, etc. The Landolt C, a circle of 5 minutes diameter with a break of 1 minute (equal to the width of the line of the character) is used also as a test character. The break can be turned up, down, etc., to test its recognition.

4.5.4.1 Different letters have different thresholds for recognition and the few letters of about equal difficulty restricts chart construction and explains why different charts give slightly different results. The differences are not great enough to be of concern in ordinary clinical practice, but can be important in research work.

4.5.4.2 Visual acuity for moving objects is different from that measured with static tests and is called dynamic visual acuity (DVA) to distinguish it from ordinary or static visual acuity (SVA). Acuity varies with the contrast of the test target and illumination, Figure 4.3. Contrast is expressed as the difference between the luminance of the object and the luminance of its surround divided by the luminance of surround. At any given intensity there is a minimum contrast which is visible. Some relations between contrast and illumination are shown on Figure 4.3.

4.5.5 Lighting, comfort, and glare. For a given intensity of illumination, contrast, and size of object, there is also a minimum time for vision. At any luminance level less time is required to see at higher luminance levels. The time relations are different for scotopic vision at low luminance levels than for photopic vision. Except on rapidly moving vehicles the time factor is usually too small during daylight to limit vision. However, in the present jet age the seeing reaction time of an individual is too great to avoid collision at the distances at which very rapidly moving aircraft can be seen.

4.5.5.1 Adequate lighting is necessary for comfortable seeing. Too little light is inadequate, leads to strain and fatigue, and with too much light (sunlight on snow or ice) temporary blindness occurs. Outdoors, the eye can see well in the shade with 100 to 400 ft-L brightness. Indoors, considerably less luminance is available (6-20 ft-L). Because of the adaptation of the eye, the indoor room appears bright at night. The amount of illumination required for seeing depends on the size and reflectivity (contrast) of the object. Sewing with black thread on black cloth requires many times the illumination needed for black thread on white cloth. Lighting recommendations of the Illuminating Engineering Society are available and a recent revision considers contrast and time for adequate vision.

4.5.5.2 Light reaching the retina other than in a useful image is called glare. Glare reduces vision most when the glare source is close to the object or is between the object and the viewer. Small amounts of glare make seeing difficult and are uncomfortable. Excessive glare disturbs the adaptive state of the eye, can prevent seeing and should be avoided. Methods for measurement and computation of glare effects are available.

4.5.6 Color vision. Color vision depends on the spectral distribution of the illumination and the wavelength range reflected or transmitted to the eye, the state of adaptation of the eye and the part of the retina involved. For example, a red object would reflect wavelengths greater than 640mμ, a blue object from 410 to 480mμ. A monochromatic yellow light (589mμ) from a sodium lamp falling on a blue object could not be reflected and the object would appear dark. A yellow can also include yellow, orange and red light. Subtractive color appears when parts of the spectrum are removed; additive color when more than one color is combined, as by projecting onto a screen. The brightness of colors depends on the energy in the light and the sensitivity of the eye, Figure 4.8. The spectral distribution of energy from different sources can be quite different, e.g. ordinary tungsten lamps are deficient in blue and produce an excess of red light as compared with sunlight. The term daylight is meaningless unless specified with respect to, time, place and direction. Average noon sunlight is nearly an equal energy spectrum, but light from a north sky has an excess of blue and a higher color temperature than direct sunlight. To avoid these ambiguities in color measurement, standard sources have been defined and internationally accepted, and any work on color vision or color comparisons should be made with standardized conditions.

4.5.6.1 The normal human eye can match any color with a mixture of three primary colors: red, green, and blue. Color blindness, that is having only gray visual sensations, is extremely rare in humans and only a few such people (achromats) have been measured and described. More common is the condition of deficient color vision, and one in ten men and one in one hundred women have more or less color vision deficiency. The most common deficiency is poor red-green discrimination, and relatively rare are defects in blue-yellow vision. A mild deficiency, or anomalous color vision, is indicated when the person requires more or less green than red to match a standard yellow, but still must have all three primaries for color matching. When the deficiency is in green, the individual is said to be deuteranomalous; when the deficiency is in the red, protanomalous. A more severe type of color deficiency is dichromatic vision. The dichromat can match any color with only two primaries. Green deficient dichromats are called deuteranopes, and the red deficient dichromats are protanopes.

4.5.6.2 The color deficient individual is unable to distinguish certain colors, and the type of color confusion points to the kind of anomaly. There are appropriate tests to determine color deficiency and such tests must be done under proper illumination. A protan who is red deficient would see red, brown, dull green, and bluish green as the same color when they have the same brightness. A green deficient deutan would confuse purplish red, brown, olive, and a green. A tritan, the rare yellow-blue deficiency, would be unable to distinguish a purple from a tan or a yellow.

4.5.6.3 Color vision may improve and reach maximum towards the end of adolescence. Thereafter, there is little change until old age. Color defectiveness is inherited and no cure or remedy is known. A mild deficiency is only a small handicap and may not even be known by the person. Medium deficiency would exclude a person from working where medium color discrimination is important, and seriously deficient individuals should be excluded from all occupations where color recognition is important. Color codes should use colors which have a minimum confusion. A good example is a green traffic light with enough added blue that it is ordinarily not confused with the red light by most color defective people. The seeing of colors is more difficult when they are small and thereby require excellent color vision ability.

4.5.6.4 The very center of the retina is color deficient for yellow. A yellow object, sufficiently far away that its image is small enough to fall in this region, appears light grey or white. Yellow has not been a very satisfactory color for air-sea rescue, because of its confusion with the white caps on the ocean. The most conspicuous color depends on the background against which it is seen and the color vision of the observer. A golden yellow, or orange is usually readily seen. Reds appear dark and may not be seen by protans.

4.5.6.5 Looking at a colored object through a complementary colored filter makes the object appear dark; conversely, through a filter of the same color it may not be seen at all. Colored glasses reduces the overall amount of light to the eye, and vision is reduced in proportion to the loss of light. With the rare exception when complementary color contrast can be used, and there is sufficient light, colored glasses will reduce seeing. This reduction is increased as dusk approaches, and no colored glass improves seeing at night. A neutral glass can reduce the intensity of light and, if not too dark, maintain color discrimination.

4.5.6.6 The appearance of many colors will change with changes in the viewing conditions. Increasing, or decreasing, the intensity of light will de-saturate some colors, and change others to a different hue. As dusk falls, a lemon yellow gradually changes to light grey or white and may not be distinguishable from a white object. For the normal eye, red is seen as red when seen as a color, but other dim colors may not be recognizable. Some colors also change in hue after being fixated for some time.

4.5.7 Perception. Perception has been defined as a complex appearing in the field of consciousness and made of sense impressions supplemented by memory. Outside of experimental projects most seeing is done at the perceptual level. The recognition of objects depends on their form and shape, and is supplemented by learning or training. It is also possible to make psychological scales, as it is possible to adjust two lights so that one appears to be twice, or half as bright as the other. The scale of equal steps in brightness can then be related to the energies measured as photometric luminances. A brightness scale increases at an exponential rate with respect to the stimulating energy.

4.5.7.1 The appearance of objects depends on their immediate surrounds, due to retinal irradiation. A series of discs cut out of the same grey paper, but placed on brighter or darker greys will not appear to be the same, but lighter or darker depending on the contrast with the surround. The appearance of color depends on the surround and on the immediately previous color adaptation. White paper looks white in daylight and will also look white at night under tungsten illumination, even though the tungsten light has more red and yellow, and the paper is reflecting more red and yellow to the eye, as the eye has adapted to and interprets the new illumination. After exposure to an intense stimulation there is seen a series of after-images. These will be in complementary color when the object is colored and they are seen against a neutral background. The after-images gradually fade and may or may not affect seeing, depending on their intensity.

4.5.7.2 Much work, during and following World War II, has discovered better form, size and arrangement for visual displays to aid the designer when scales or indicators are needed. Vision through instruments involves the same principles discussed in this section. Unless the instrument produces a sharp image of proper size, intensity, and contrast on the retina it cannot be resolved and seen. Glare should be avoided. Reticles and scales that appear in the field of view require careful planning as to size, contrast, and lighting if they are to be seen with comfort. When half shade plates, or comparison fields are used in an optical instrument, the dividing lines should become invisible and the areas compared should have the same size, otherwise a slightly larger lighter area may be equated with a slightly smaller darker area.

4.6 MOVEMENT OF THE EYES

4.6.1 General. Six muscles move the eye. The conjunctiva, Tenon's capsule, and the fat pads within the orbital cavity of the skull aid in positioning the moving eye. The center of rotation is about 13-15.5 millimeters behind the cornea. Since there are no inflexible mechanical axes, the center of rotation may vary a millimeter or so depending on the resultant of the muscular action. The muscles which turn the eye are coordinated with those of the other eye, by the muscular movements within the eye, by the movements of the eye lids, and also by the neck muscles which move the head via the nervous system.

4.6.2 Muscular action. The superior and inferior rectus muscles as illustrated in Figure 4.10, raise and lower the eye in a plane 23° from the plane of the medial orbital wall. This is the wall of the skull seperating the nasal and orbital (eye) cavities. The medial (internal) and lateral rectus muscles rotate the eye toward or away from the nose in a horizontal plane, when the eye is in the primary position of looking straight ahead. The superior oblique muscle passes through tendon pulley and inserts into the upper, back side of the eye so that contraction of the muscle depresses the eye. The inferior oblique muscle is attached underneath the eye and on contraction raises the eye. The movement of the oblique muscles is in a plane through the center of rotation of the eye which slopes back about 129° from the medial orbital plane. The gaze must be directable to any place within its field of view, Figure 4.11, and maintain a horizontal reference on the retina corresponding with horizontal in the field of view. The superior oblique and the inferior rectus muscles working together

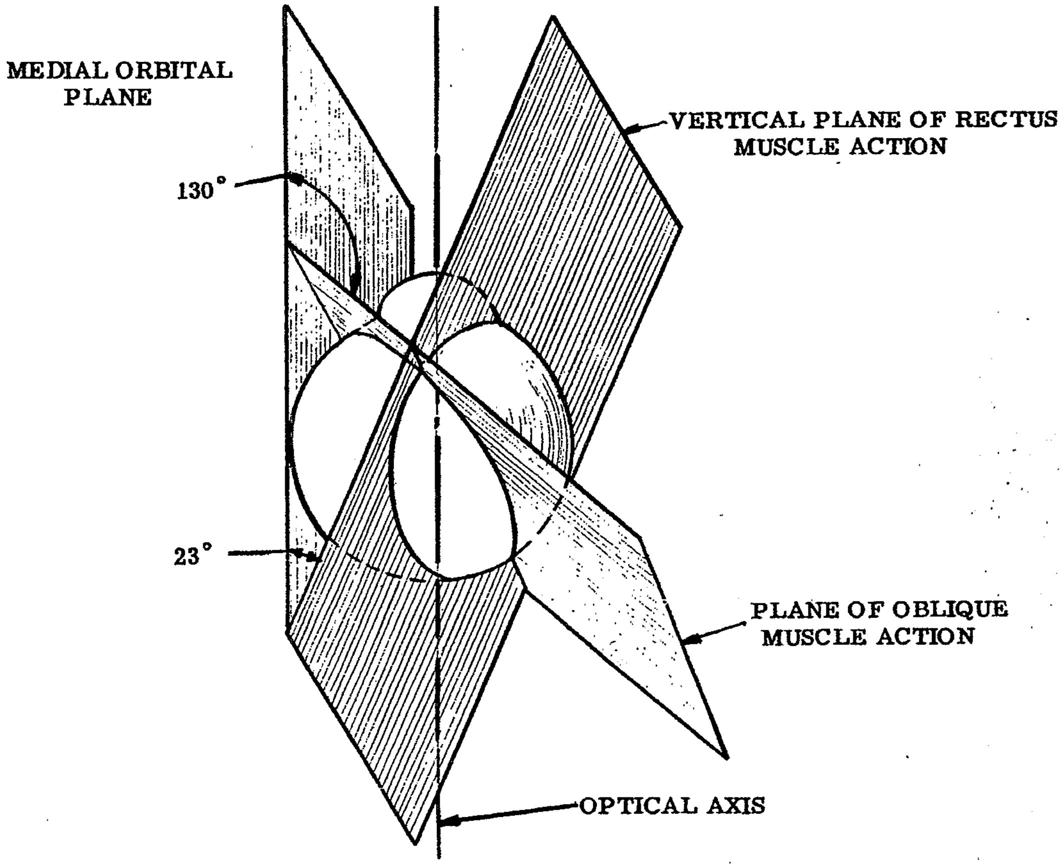

MEDIAL ORBITAL PLANE
VERTICAL PLANE OF RECTUS MUSCLE ACTION
130°
23°
PLANE OF OBLIQUE MUSCLE ACTION
OPTICAL AXIS

Figure 4. 10 - Planes of rotation of the external eye muscles.

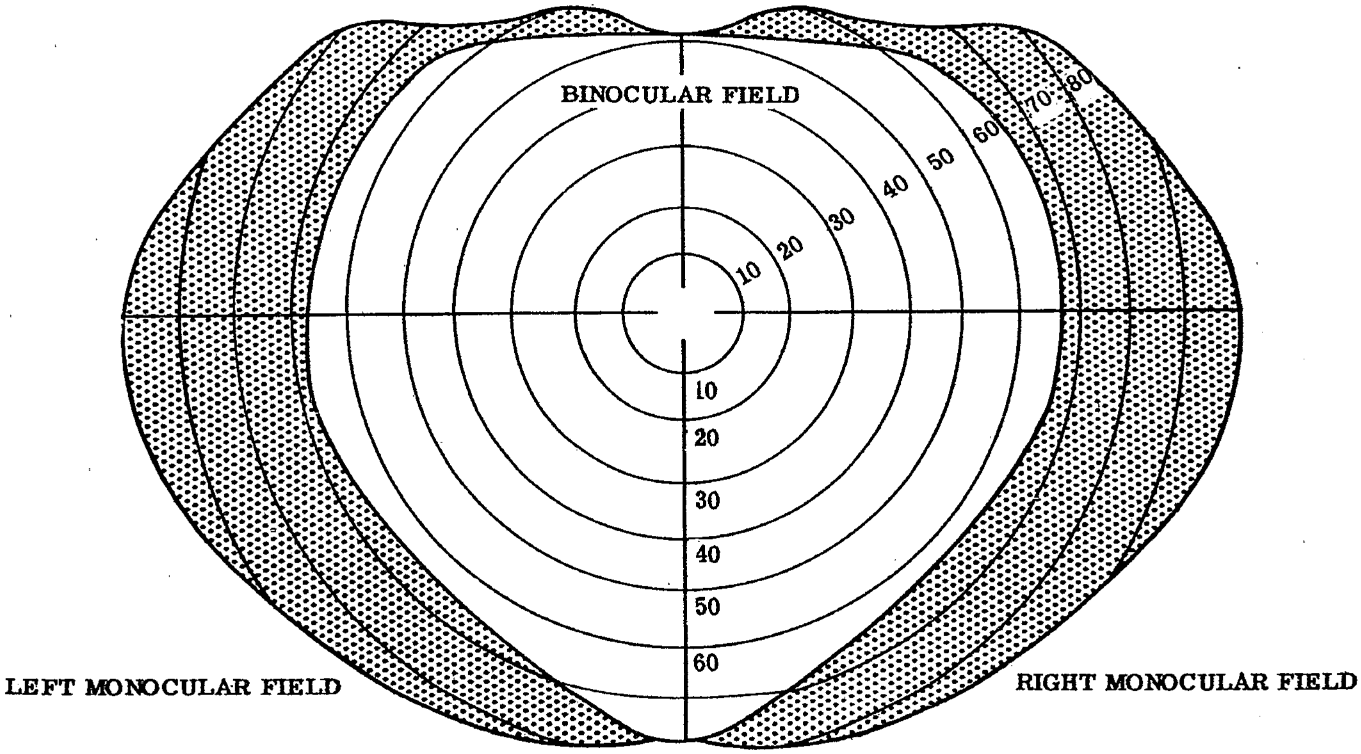


Figure 4. 11. - Monocular and binocular visual fields.

minimize a tendency for the eye to roll on an anterior posterior axis. Nevertheless, there is some torsion, or rolling, of the eye that can be mapped with the aid of after-images. Plotting the observations shows the visual field to have pincushion distortion.

4.6.3 Imbalance. The actual motions of the eye are complex. Adjustments of the eye to the left or right are made easily and up or down reasonably well, but the eye muscles are not arranged for movement of the eyes at oblique axes. Consequently, if the eyes are provided with more than slightly twisted images, they cannot adjust and fuse for single vision. The movement of the eye from one fixation to another is not smooth, and consists of movements of about 4 minutes subtended arc. The eye does not move directly to the point of fixation, but moves towards it and then approaches the fixation by a series of smaller movements. The following of a moving object by the eye also tends to go in small jumps rather than as a single smooth movement. The movements are the resultant of the contractions of one or more pairs of opposed muscles, and fluctuations are characteristic of neuromuscular mechanisms. Action potentials of the muscles can be recorded from electrodes placed around the eye, or within the muscles and their analysis is providing considerable new information on muscular movements. The eye follows a moving object as far as it can and then suddenly jumps back to a new fixation and this stepwise motion is called physiological nystagmus. Workers in mines under dim light develop a characteristic nystagmus.

4.6.4 Phorias and trophias. Two types of misalinement of the eyes have clinical importance. When one eye is covered and subsequently moves away from the fixation point, the condition is called a phoria. If the visual axes of the eyes are different when the eyes are open and uncovered, the condition is called strabismus or squint, and the direction is indicated by a tropia. Normal fixation is orthophoria or orthotropia and deviations would be heterophoria or heterotropia. The direction of the abnormal orientation is indicated by prefixes: eso- refers to movement toward the nose, exo- toward the temple, cyclo- a rotation, hyper- up, and hypo- down. Esotropia would indicate crossed eyes, while esophoria would indicate a moving toward the nose by a covered eye, or when the eye is dissociated from binocular vision.

4.7 BINOCULAR VISION

4.7.1 Advantages. The use of two eyes is a decided advantage in seeing. There is an apparent increase in brightness of about 20% when an object is seen with both eyes rather than with one eye alone. Normally the eye movements are equal and symmetrical and the sensory feed-back from the movements aids in balance and orientation of the organism.

4.7.2 Stereoscopy. A great advantage of two eye vision is the emergence of the experience of depth, or stereoscopic vision. Stereoscopic depth is a primary factor. Other factors which aid in the understanding of depth, such as superposition, are learned secondary factors. The basis of stereoscopic vision is horizontal dissimilarity of retinal images on corresponding points of the two retinas. In Figure 4.5, looking at the two points A and B which are at different distances from the eye, the images of the lines at A and B for the left eye are closer together than for the right eye. The fusion of these dissimilar images leads to the space perception that one is farther away from the other. Likewise, if one arranges drawings to give disparate images (within the physiological limits of the eye) when viewed through a stereoscope, the appearance of depth is produced. Stereopsis varies with the distance between the centers of the two eyes, the interpupillary distance (PD), and the spacing of the eyes alters the spatial visual geometry.

4.7.2.1 In designing binocular instruments, sufficient adjustment must be provided for the interpupillary distance of the intended observers. Formerly, 50 to 75 millimeters was considered adequate, but individuals are now growing larger and 76 millimeters maximum interpupillary distance have been used.

4.7.2.2 In stereoscopic depth the disparity between the retinal images for contours is probably more important than mere difference in size. There are limited areas on the retinas, within which objects can be placed on corresponding parts of the retinas, called Panum's areas. These areas are probably accounted for by the extent of the overlapping of the arborizations of the neurones from corresponding retinal points at the terminal areas of the cortex of the brain. The stereoscopic threshold is the smallest depth or disparity that can be experienced, and depends on the dimensions, contrast sensitivity of the retinal elements, and the sharpness of focus, i.e. the size of the blur circle on the retina. Stereoscopic acuity is less for individuals with less than 20/20 vision, but fails to increase with superior visual acuity. Stereoscopic vision is not limited to the macula and there is some evidence that it is maximal at an extra-foveal angle of 15-21 minutes. Useful stereoscopic depth is limited to about 1900 feet or a disparity angle of 24 seconds. For stereoscopic range finders the unit is about 12 seconds. The threshold for stereoscopic perception of depth increases with decreased illumination in dark adaptation, and shows a marked change which corresponds with the shift from photopic to scotopic vision.

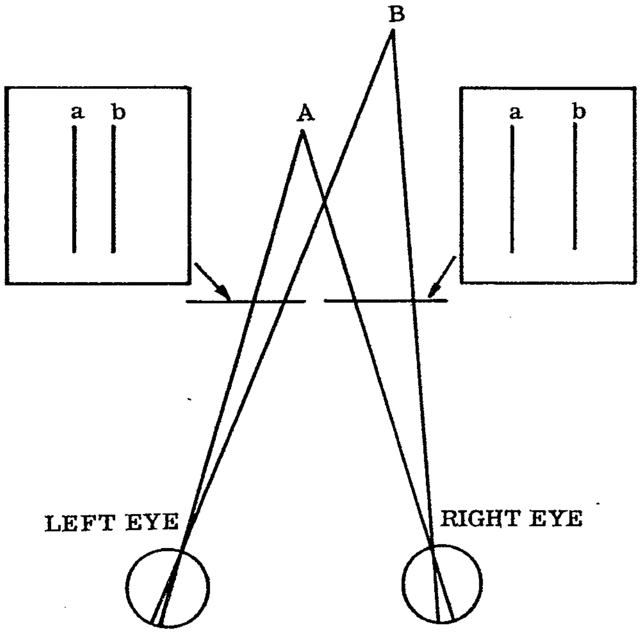


Figure 4. 12. - Stereoscopic vision with disparate images.
(From American Medical Association, Archives of Opthalmology, No. 60, K. N. Ogle, 1958)

4.7.2.3 One of the main problems in vision is the interpretation of the geometry of what we see. This involves the two eyes, their separation, and the connections within the brain. If we use a neutral filter to absorb some of the light to one eye, little change is noted in the stereoscopic effect for static objects, but if we look at a pendulum we find that the apparent movement is no longer in a single plane, but the bob tends to swing around an ellipse. This Pulfrich illusion is explained as a result of the different reaction times for the eye with and without the filter.

4.7.3 Psychological and physical space variations. Psychological visual space is different from Euclidean physical space. If five lights are arranged in a dark room to be in a straight line they will be found to be in a curved line after the lights are turned on. When aligned at right angles to straight ahead gaze, one plane is found where the lights would be set in a straight line. Nearer than that, the lights would be in an arc concave toward the eye and farther away in an arc convex to the eye. Such experiments provide evidence that psychological visual space is hyperbolic or elliptical rather than Euclidean. The transformation equations between physical and psychological space have not been fully worked out.

4.7.4 Limitations. There are practical applications for instrument design. If the images of an object are different in each eye either a depth sensation or distorted space perception will occur. When the differences are due to unequal magnification in size the appearance is that of a distorted space, and space distortion from size differences in the images is aniseikonia. The tolerance of individuals to such differences varies, but differences of 1 to 2% or more usually result in visual strain and discomfort. Differences of 5% usually preclude binocular vision. The differences are not always those of the actual size of the images on the retina but rather are an overall size effect which involves the central nervous system. An Eikonometer is used for clinical measurement and the aniseikonia can be corrected by a special size lens for one eye. Differences in size are innate in some eyes. In others they are produced artificially by a considerable difference in the spectacle prescription for the two eyes. A common problem arises from unilateral aphakia, when a strong, plus-spectacle lens is needed to take the place of the lens of the eye. It may not be possible under these conditions to restore stereoscopic binocular vision.

4.7.5 Design considerations. The design of binocular instruments is challenging since comfortable viewing with two eyes presents difficulties that do not occur with monocular instruments. The coordinated motion of the two eyes must not be disturbed. A pupillary adjustment of 50 to at least 76 millimeters should be provided. Magnification differences to the two eyes should not exceed 2%. Some people cannot tolerate more than 0.5% while others may tolerate a little more than 2%. Oculars must be paired so that increased size differences will not occur. Beam splitters should be neutral, otherwise the light to the two eyes will cause discomfort from the chromatic aberration of the eye. Should one eye receive a bluish light and the other eye a redish light the accommodation of each would have to be different, which would lead to strain and intolerable discomfort. The amount of light to the two eyes should be balanced, preferably within 10%. Vertical imbalance should not exceed 0.5 prism diopter. Horizontal imbalance need not be quite so small, but in excess of this value it would be fatiguing. Spectacle prescription practice holds to about 0.25 prism diopter. For low power instruments such as a bi-objective, binocular, microscope a 0.33 prism diopter difference may be tolerable. Any twist in the images should be kept to a minimum to avoid strain from complicated and difficult eye movements necessary to aline the images on the retinas. Since the light is divided to two eyes, more light will be required for binocular than for monocular instruments. In some types of binocular instruments, double mirrors or a large or diffusing mirror, may be necessary to direct the light to both eyes. When the objectives and the oculars of the instrument have different convergent angles, the appearance of depth can be made true (orthoscopic), or it can be increased or decreased (hyper- or hypostereoscopy), providing another variable for use by the instrument designer.

4.8 FATIGUE AND AGEING

4.8.1 Fatigue. Fatigue of the retinal processes is not likely at ordinary conditions. The usual "visual fatigue" (asthenopia) is muscular rather than retinal. Difficult seeing gradually involves 20 or more muscles, spreading to include those of the brow, cheek and lip. Greater mental effort is needed for getting and interpreting the visual information required. Uneven lighting results in one part of the retina needing more light and calling for pupil opening, while another is over stimulated and calling for a smaller pupil. The resultant conflict fatigues the ciliary process. Changes in illumination, too rapid for the accommodating ability of the eye, cause local and general fatigue. Continuous use of more than one-half of the available accommodative response, and close work necessitating strong convergence are fatiguing. Body tension increases during difficult seeing. An awareness of body sensation during difficult seeing, and the appearance of increasing hyper-reactivity, both increase general fatigue. A visual perceptual load, greater than can be assimilated, is also fatiguing. Visual fatigue is minimized with proper illumination, adequate contrast, form and time for seeing, proper arrangement for easy functioning of the eyes, and comfortable working conditions. An uncomfortable posture can cause eye strain and fatigue especially if seeing becomes difficult (dim light, fog, glare, etc.). An unpleasant task may make the eyes feel very tired, although instant recovery may occur on changing to an interesting visual task.

4.8.1.1 Any instrument that requires steady orientation of the eyes should be provided with a head rest, and heavy equipment should be properly supported in order to lessen fatigue. Instruments should be set up so that they are observed with a straight ahead position of the eyes, and when that is not feasible the instrument should be adjusted to the head for comfortable vision, not the whole body of the observer cramped into a viewing position. Image brightness and convergence should be adjustable and no adjustments of the eyes beyond normal functional ability should be required by an optical instrument (unless designed to test a visual function).

4.8.2 Age. Seeing is probably at its best towards the end of adolescence. Some of the age changes are summarized in Figure 4.13. At about age 40 the accommodative mechanism begins to fail and the individual is no longer able to focus the eye on near objects. This is due to a decrease in the elasticity of the lens of the eye, although the focussing muscles may also be involved. The condition is called presbyopia and is corrected by adding positive spherical power to the spectacles, usually in the form of a bifocal, or trifocal addition. The trifocal addition has the further advantage of providing an intermediate distance of clear vision just beyond that of the near correction. The pupil of the eye does not open as far in the elderly, which fortunately increases the depth of field. Although less light gets to the retina and greater illumination is necessary for equal visual efficiency. One experimenter has found that the illumination should be doubled for each 13 years increase in age.

4.8.2.1 The eye media lose transparency, particularly the lens, which becomes yellowish as age increases. These changes effect color vision, and in addition, lessen the light available for image formation. Accommodation is slower in old age than in youth. The efficiency of the retina declines and resistance to glare becomes less. The fibers of the lens may become opaque and form a cataract. With developing cataracts, asymmetrical screening may improve the vision slightly by reducing glare. The balance between enough light for adequate seeing, and excess light or glare, is difficult and more critical in later life. When instruments are to be designed for use both by young and old people the limitations of the older eye should be kept in mind.

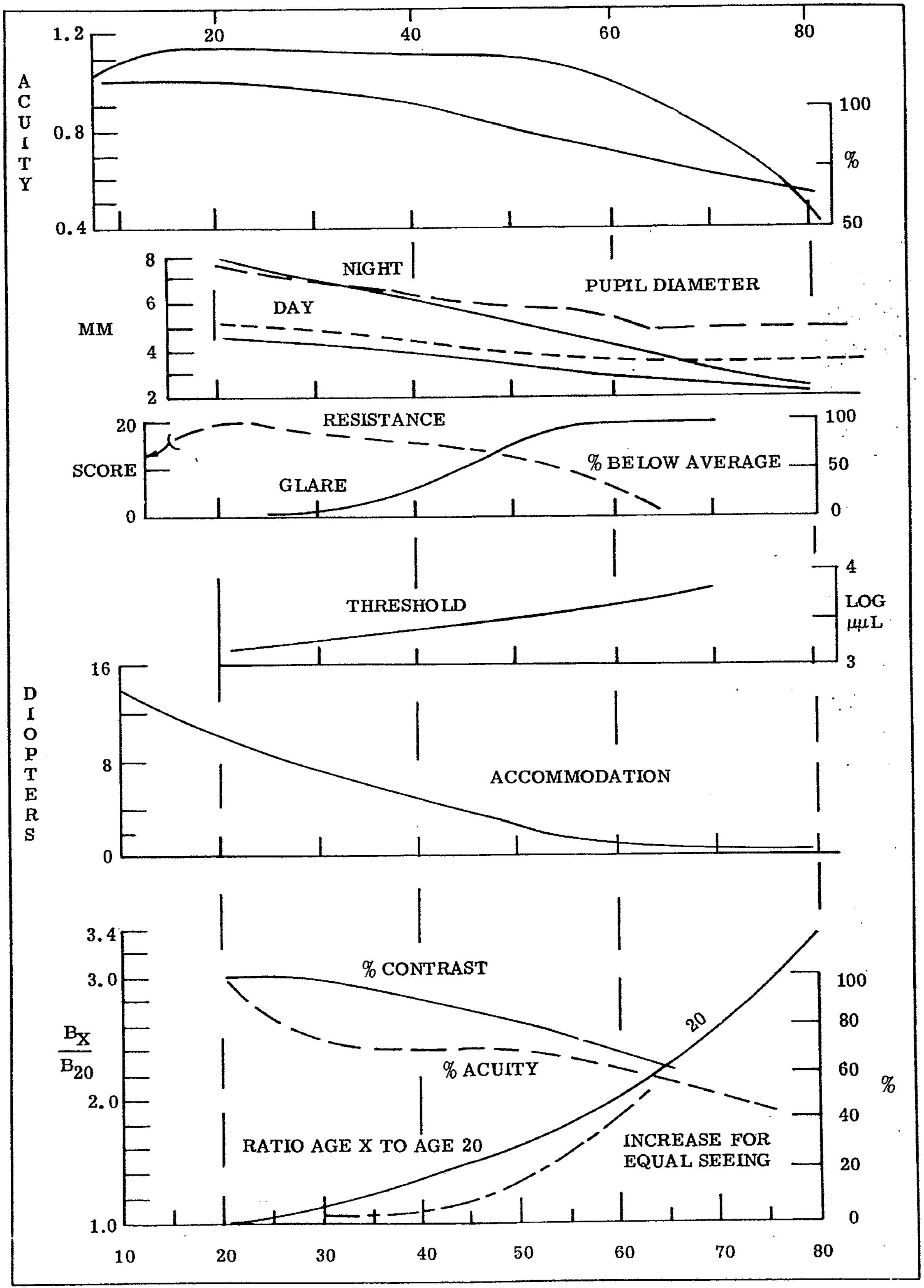


Figure 4.13 - Some age changes in vision.
(From American Journal of Optometry, O.W. Richards, 1958)

5 FUNDAMENTAL METHODS OF RAY TRACING

5.1 GENERAL

5.1.1 Basic optical system. Every optical system consists of one or more reflecting or refracting surfaces. The function of the system is to transform the diverging spherical wavefronts coming from object points in object space to converging spherical wavefronts going towards image points in image space. As mentioned in paragraph 2.1.2 the passage of the wavefronts through the optical system can be most easily discussed by utilizing the concept of rays. The passage of rays through an optical system may be determined by purely geometrical considerations, since it is correct to make the following assumptions:

(1) A ray travels in a straight line in a homogeneous medium.

(2) A ray reflected at an interface obeys the law of reflection.

(3) A ray refracted at an interface obeys the law of refraction.

Computing the passage of rays through an optical system is a purely geometric problem best solved by the techniques of analytic geometry.

5.1.2 Centered optical systems.

5.1.2.1 Fortunately, nearly every theoretical optical system consists of centered refracting or reflecting surfaces. In a centered optical system all surfaces are rotationally symmetrical about a single axis. A cross-section view of a typical photographic lens is shown in Figure 5.1. In this case all the surfaces are spherical surfaces and the centers are assumed to lie on the optical axis. Herein lies one of the differences between theory and practice. In the design phase, the system is assumed to have an axis of symmetry. In practice the lenses may not be lined up perfectly so it will not be a centered optical system. If the lens is to perform according to the design, the lenses must be adjusted until they are centered. Procedures to assure centering of the elements are a prime consideration in the mechanical design of optical instruments.

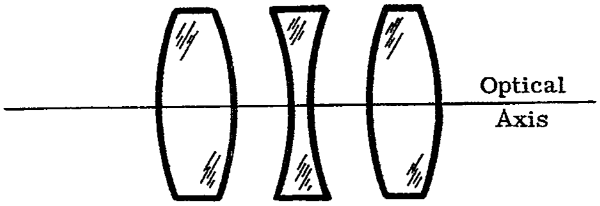


Figure 5.1 - A cross-section view of a photographic lens.

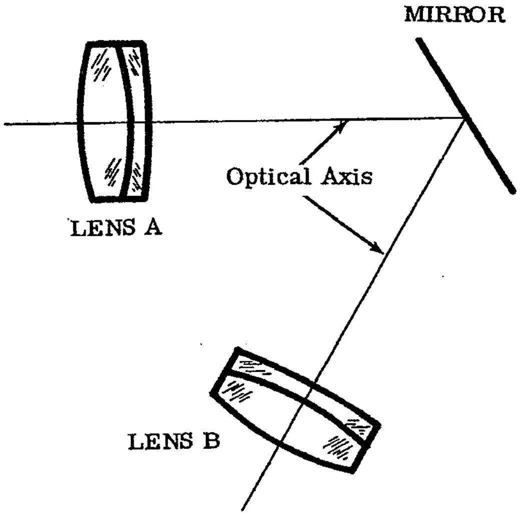


Figure 5.2 - An optical system containing a mirror.

5.1.2.2 The optical system shown in Figure 5.2 may not appear at first glance to be a centered optical system. The optical axes of the two lenses do not coincide. However, if properly constructed this may be a centered optical system. To understand this consider Figure 5.3, which shows how a system involving plane mirrors can be thought of as folded out. These ideas are treated in detail in Section 13.

5.1.2.3 Consideration of the law of reflection shows that the ray of light traveling along the optical axis from lens A is actually deflected, but can be thought to continue straight through the mirror. If the axis of lens B lies on the extended axis of lens A, then the system is a centered optical system. One can see that if lens B of Figure 5.3 is shifted to the left or right, there will be a corresponding shifting of lens B' up or down; the system will become decentered and lose its axial symmetry.

5.1.2.4 The sections on geometrical optics in this handbook consider centered systems. Decentered systems usually, when carefully analyzed, are seen to be part of some centered system. Hence if a final design calls for a decentered system, the preliminary design considers the centered system as a basic starting point.

5.1.3 Plane, spherical and aspheric surfaces.

5.1.3.1 Production techniques for generating plane and spherical surfaces on optical materials are well established and thus these are most commonly used. Aspheric surfaces, however, offer certain advantages, and recent advances in the generation of this type of surface, coupled with the need for the design refinements they offer, have resulted in more frequent design application of this type. Aspheric surfaces are also usually considered to have rotational symmetry about the optical axis.

5.1.3.2 In ray tracing, plane surfaces will be considered to be special cases of spherical surfaces, having radii equal to infinity; hence no special technique for plane surfaces will be developed in detail in this section. In Section 13, reflection from plane surfaces is considered more fully. The technique for treating aspheric surfaces is developed by extending the technique for spherical surfaces. In both cases, the surfaces are considered to be centered.

5.1.4 Ray tracing, the basic tool of optical design.

5.1.4.1 In order to understand clearly the kind of image formed by a system, and what must be done to im-

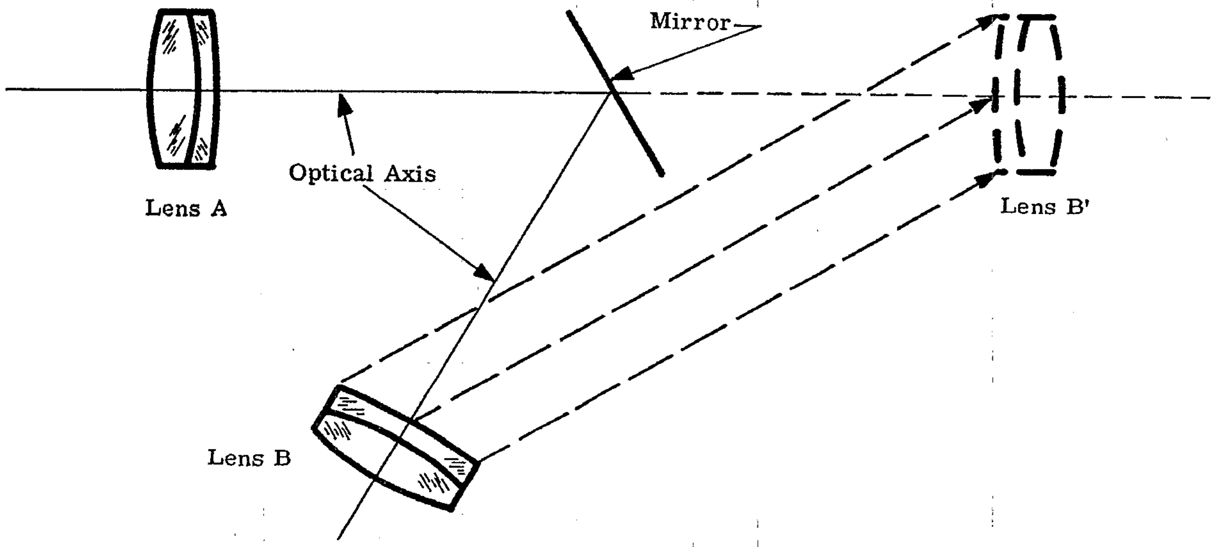


Figure 5.3 - Diagram showing "folding out" of an optical system containing a mirror.

prove this image, a certain number of rays must be determined in their passage through the system. This process of ray tracing involves the determination of the direction and location in space of each segment of a ray as it goes from object to image. Since the function of the system is to transfer light from an object surface to an image surface, the object surface and the image surface, although neither reflecting nor refracting, can be considered as surfaces of the optical system.

5.1.4.2 Figure 5.4 shows a cross-section view of a centered optical system. The ray, consisting of seven straight line segments, goes from the object point, O, on the object surface, to the image point, O', on the image surface, being refracted at six intermediate surfaces. The remainder of Section 5 will be concerned with numerical and graphical methods of determining the course of general and special rays through a general system.

5.2 DEFINITIONS AND CONVENTIONS

5.2.1 Need for specific conventions. The ray tracing formulae to be used for tracing a ray through a system involve parameters of more than a single surface or a single medium. Therefore, it is important to adopt a convention of notation which will clearly distinguish one surface from another and one medium from another. In addition, many optical systems employ mirrors, so that the rays sometimes proceed in a direction generally opposite to the incident rays. Our conventions should be such that a reflecting surface can be handled as any other general refracting surface. It is assumed that before applying these conventions the system has been folded out in the sense of Figure 5.3.

5.2.2 Statements of definitions and conventions. The following definitions and conventions, which are in agreement with those given in MIL-STD-34, will be used in Sections 2, and 5 through 15, inclusive. Reference to Figures 5.4 and 5.5 will indicate examples of some of these conventions.

(1) It will be assumed that light initially travels from left to right.

(2) An optical system will be regarded as a series of surfaces starting with an object surface and ending with an image surface. The surfaces will be numbered consecutively, in the order in which the light is incident on them, starting with zero for the object surface and ending with k for the image surface. A general surface will be called the j th surface.

(3) All quantities between surfaces will be given the number of the immediately preceding surface.

(4) A primed superscript will be used to denote quantities after refraction only when necessary.

(5) r_j is the radius of the jth surface. It will be considered positive when the center of curvature lies to the right of the surface.

(6) The curvature of the jth surface is $c_j = 1/r_j$. c_j has the same sign as r_j.

(7) t_j is the axial thickness of the space between the jth and the j + 1 surface. It is positive if the j + 1 surface physically lies to the right of the jth surface. Otherwise it is negative.

(8) n_j is the index of the material between the jth and the j + 1 surface. It is positive if the physical ray is traveling from left to right. Otherwise it is negative.

(9) K_j, L_j, M_j are the products of n_j and the direction cosines (with respect to the X, Y, Z axes respectively) of a ray in the space between the jth and the j + 1 surface. They will be called the optical direction cosines.

(10) The right-handed coordinate system shown in Figure 5.5 will be used. The optical axis will coincide with the Z axis. The light travels initially toward larger values of Z. Positive values of X are away from the reader in Figure 5.5.

(11) X_j, Y_j, Z_j are the position coordinates of a ray where it intersects the jth surface.

(12) In writing formulae where no confusion is likely to result, the j will be omitted from the subscript. Thus the curvature of the j - 1 surface will be written c_{-1}, the curvature of the j th surface will be written c and the curvature of the j + 1 surface will be written c_{+1}.

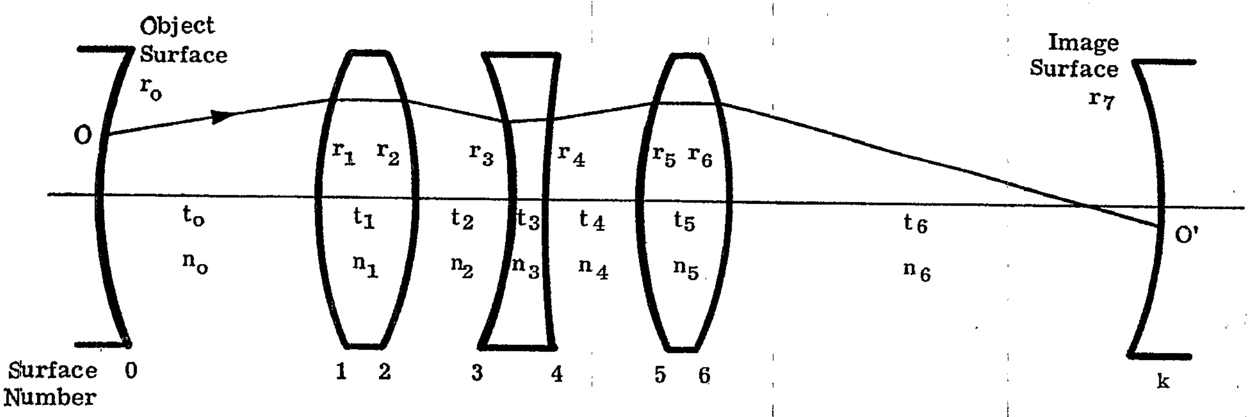


Figure 5. 4–Cross-sectional view of a centered optical system.

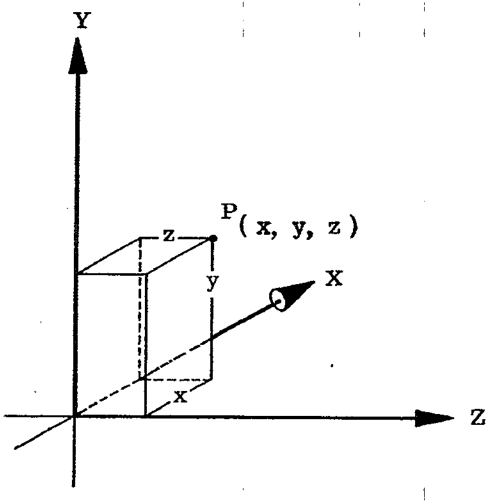


Figure 5. 5- Right-handed coordinate axes.

5.3 BASIC RAY TRACE PROCEDURE

5.3.1 Transfer procedure. As can be seen from Figure 5.4 a ray travels in a straight line from a point on one surface to a point on the following surface. It is then refracted and proceeds to the next surface in a straight line. The ray tracing procedure then consists of two parts, the transfer procedure, and the refraction procedure. The transfer procedure involves computing the intersection point of the ray on the surface from the optical direction cosines and the intersection point data at the previous surface. That is, given K_{-1}, M_{-1}, L_{-1} and X_{-1}, Y_{-1}, Z_{-1}, compute X, Y, Z. The equations used are called the transfer equations.

5.3.2 Refraction procedure. The refraction procedure involves computing the optical direction cosines of a ray from the intersection point data and the optical direction cosines of the previous ray segment. That is, given X, Y, Z and K_{-1}, M_{-1}, L_{-1}, compute K, L, M. The equations used are called the refraction equations.

5.3.3 Repetition for successive surfaces. After having applied the two procedures, we have the initial data for the next application. The transfer equations will be used to compute X_{+1}, Y_{+1}, Z_{+1} and the refraction equations will be used to compute K_{+1}, L_{+1}, M_{+1}. It should be noted that it is often convenient to introduce fictitious or non-refracting surfaces to simplify the procedure. One example is the tangent plane, an XY plane tangent to a physical surface at the optical axis. Another example is a sphere, tangent to an aspheric surface at the optical axis. These fictitious surfaces are handled in exactly the same manner as a physical surface. Transfer equations are used to go to or from such a surface. The refraction equation reduces to $I = I'$, and the direction cosines of the refracted ray equal those of the incident ray, as would be expected at a non-refracting surface. Fictitious surfaces will be used in the next section.

5.4 SKEW RAY TRACE EQUATIONS FOR SPHERICAL SURFACES

5.4.1 Types of rays.

5.4.1.1 A general ray is any ray passing from any object point through the optical system to its image point on the image surface. A special ray that lies in a plane containing the optical axis and the object point is called a meridional ray. Any non-meridional ray is a skew ray. A ray close to the optical axis is a paraxial ray. Because of the approximation involved, a paraxial ray is a special type of meridional ray. A skew ray is considered to be non-paraxial since it is non-meridional. These distinctions will become apparent as the subject is developed.

5.4.1.2 Corresponding to the three types of rays, skew, meridional, and paraxial, we will develop three sets of ray trace equations and procedures. Because the three rays, in the order given here, become less general and more specialized, the equations relating to these types of rays become simpler as we proceed from skew through meridional to paraxial. One method of developing the subject would be to discuss the simplest case first (paraxial), then proceed to the more complicated (meridional) and finally to the most general (skew). This procedure would have the advantage of beginning with the simplest derivation. However, it would necessitate three separate derivations.

5.4.1.3 We will proceed in the other direction, beginning with the most general case, the skew ray trace. From this the meridional and paraxial equations follow by simplification; hence only one derivation is necessary, instead of three. The particular equations derived in Section 5.4 are set up in a form for an electronic computer. However they are completely satisfactory for use with a desk calculator, and represent a good starting point for the human computer who has not yet worked out his own equations.

5.4.2 Initial data for a skew ray. Figure 5.6 shows the skew ray as it traverses the space between two surfaces. At the right hand surface it is refracted, and a drawing corresponding to Figure 5.6 could show this ray as it traverses the space between the j th and j_{+1} spherical surfaces. Similarly, another drawing could show the ray before refraction at the j_{-1} surface. The initial data for the ray we are considering will consist of the emergence point with the left surface, and the direction of the ray in space. Hence we specify X_{-1}, Y_{-1}, and Z_{-1}, the coordinates on the j_{-1} surface, and K_{-1}, L_{-1}, and M_{-1}, the optical direction cosines of the ray. From these data we will determine the intersection of the ray with the next surface, and the optical direction cosines of the refracted ray. These values then become the initial data for the new ray, and the process is repeated until the image point is reached.

5.4.3 Transfer procedure, physical surface to next tangent plane.

5.4.3.1 The first part of the problem, namely the determination of the intersection of the ray with the jth spherical surface, will be divided into two parts: first, the intersection of the ray with a non-physical surface, the plane tangent to the spherical surface, and, second, the final intersection with the spherical sur-

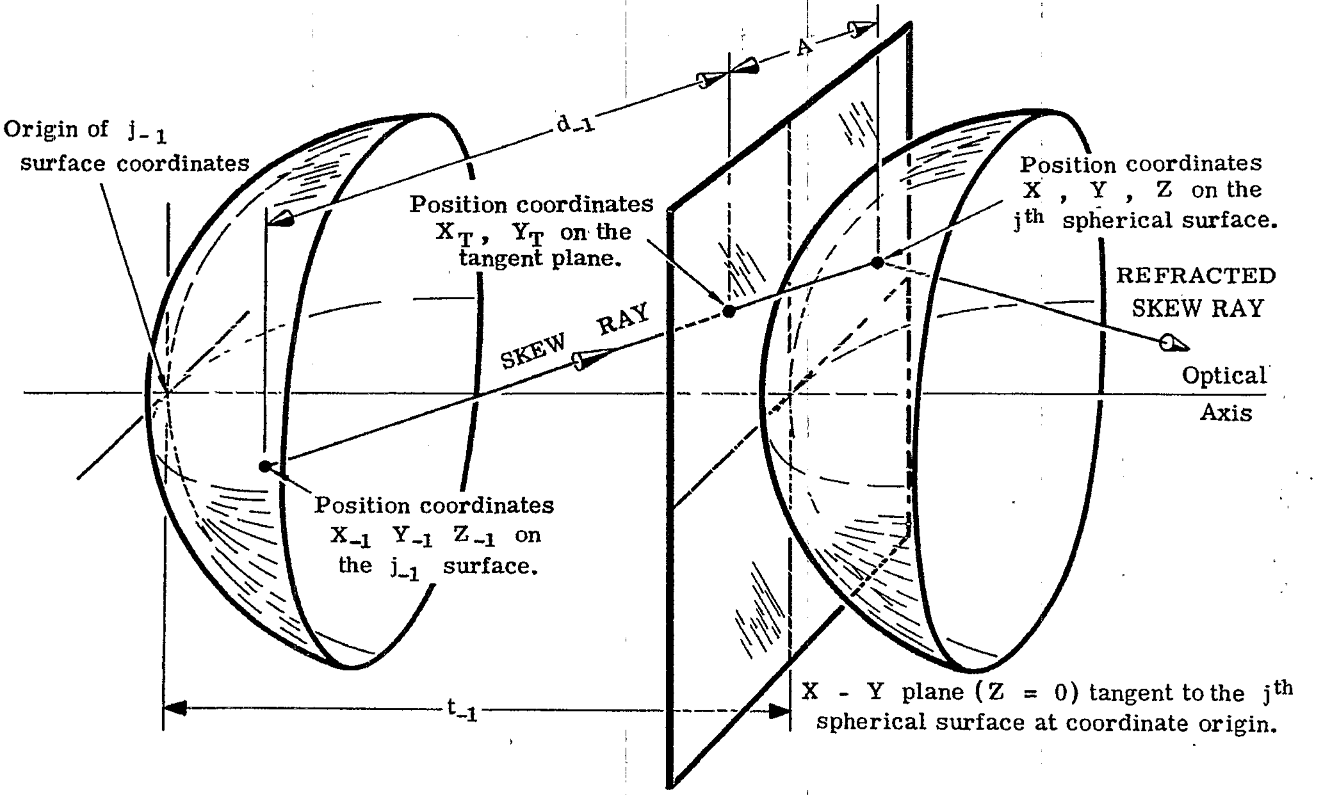

Origin of j_{-1} surface coordinates
d_{-1}
A
Position coordinates X_T, Y_T on the tangent plane.
Position coordinates X, Y, Z on the j^{th} spherical surface.
SKEW RAY
REFRACTED SKEW RAY
Optical Axis
Position coordinates X_{-1} Y_{-1} Z_{-1} on the j_{-1} surface.
t_{-1}
X - Y plane (Z = 0) tangent to the j^{th} spherical surface at coordinate origin.

Figure 5.6 - Diagram of a skew ray in space between the j_{-1} surface and the j th surface.

face. In Section 5.4.3 we consider only the first part.

5.4.3.2 The origin of the position coordinates for points on the tangent plane is at the point of tangency, the optical axis. Hence $Z_T = 0$ for all points in the plane. The new value of X, X_T, is the old value, X_{-1}, plus the change in X, ΔX. The latter is the projection of the skew ray, of length d_{-1}, onto the X axis. Hence

$$X_T = X_{-1} + \Delta X = X_{-1} + d_{-1} \frac{K_{-1}}{n_{-1}} ,$$

since K_{-1}/n_{-1} is the direction cosine of the ray with respect to the X axis. There is a corresponding equation for Y_T.

5.4.3.3 The length of the ray, d_{-1}, between the left-hand surface and the tangent plane is not given; it must be calculated from the initial data. From Figure 5.6 the change in Z is given by

$$\Delta Z = t_{-1} - Z_{-1} ,$$

and this equals the projection of the ray along the Z axis. Therefore

$$\Delta Z = d_{-1} \frac{M_{-1}}{n_{-1}} .$$

5.4.3.4 It is now possible to summarize the three equations which are used to calculate the intersection of the ray with the tangent plane.

$$\frac{d_{-1}}{n_{-1}} = (t_{-1} - Z_{-1}) \frac{1}{M_{-1}} , \qquad (1)$$

$$Y_T = Y_{-1} + \frac{d_{-1}}{n_{-1}} L_{-1} , \qquad (2)$$

and

$$X_T = X_{-1} + \frac{d_{-1}}{n_{-1}} K_{-1} . \qquad (3)$$

It should be pointed out that in addition to the initial data for the ray, we must be given the value t_{-1}, the distance between the surfaces measured along the optical axis. It is not necessary, however, to know explicitly the value of n_{-1} at this time. The specific procedure followed is first, to use Equation (1) to calculate the numerical value of d_{-1}/n_{-1} ; second, to use the value thus obtained in Equations (2) and (3) to calculate Y_T and X_T respectively.

5.4.4 Transfer procedure, tangent plane to spherical surface.

5.4.4.1 The discussion in Section 5.4.3 treated the first part of the transfer problem. The following discussion treats the second part, transferring the ray coordinates on the tangent plane to those on the spherical surface.

5.4.4.2 Referring to Figure 5.6, since the tangent plane is not a refracting plane, the ray continues on to the sphere, for a distance A. The segment A has the same optical direction cosines as the segment d_{-1}. Therefore the new values of the coordinates, X , Y , and Z on the sphere, are determined from the values on the tangent plane, X_T, Y_T, and Z_T, by the process that was used to set up Equations (2) and (3). Remembering that Z_T is zero, we have

$$X = X_T + \frac{A}{n_{-1}} K_{-1}, \tag{4}$$

$$Y = Y_T + \frac{A}{n_{-1}} L_{-1}, \tag{5}$$

and

$$Z = \frac{A}{n_{-1}} M_{-1}. \tag{6}$$

5.4.4.3 In order to use Equations (4), (5), and (6), it is necessary to calculate the value of A. It is clear from Figure 5.6 that this value depends on the curvature of the jth spherical surface, the coordinates of the ray at the tangent plane, and the direction cosines of the ray. We will use a relation between X , Y , Z and c which depends on the properties of a sphere. This equation can be used with Equations (4), (5), and (6) to eliminate X , Y , and Z . The result will be an expression for A/n_{-1} in terms of known data.

5.4.4.4 Figure 5.7 shows a plane containing the optical axis and the intersection point (X , Y , Z) of the

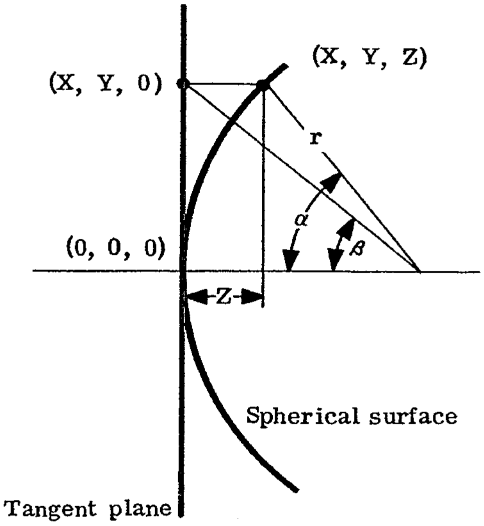


Figure 5.7 - Some properties of a spherical surface.

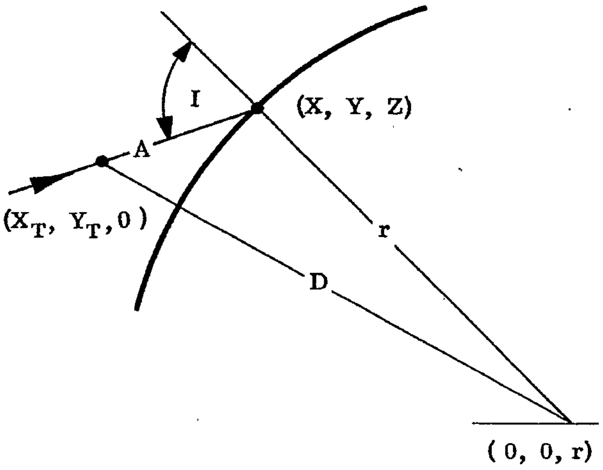


Figure 5.8 - Determination of $n_{-1} \cos I$.

ray on the spherical surface. From the figure, and recalling that $c = 1/r$, we have

$$Z = r - \left[r^2 - (X^2 + Y^2) \right]^{1/2} = \frac{1}{c} - \frac{1}{c}\left[1 - c^2 (X^2 + Y^2) \right]^{1/2},$$

which can be simplified, by transposing and squaring, to

$$c^2 (X^2 + Y^2 + Z^2) - 2cZ = 0.$$

Substituting into this equation the expressions for X, Y, and Z from Equations (4), (5), and (6). The result, on collecting terms, is

$$\left(\frac{A}{n_{-1}}\right)^2 c\,(K_{-1}^2 + L_{-1}^2 + M_{-1}^2) - 2\left(\frac{A}{n_{-1}}\right)\left[M_{-1} - c\,(Y_T L_{-1} + X_T K_{-1}) \right] + c\,(X_T^2 + Y_T^2) = 0.$$

5.4.4.5 In this last simplification it was assumed that $c \neq 0$; the case of $c \to 0$ will now be considered. Since the sum of the squares of the direction cosines is unity, the coefficient of $(A/n_{-1})^2$ is $c\,n_{-1}^2$. Calling the other coefficients $2B$ and H respectively, we have

$$c\,n_{-1}^2 \left(\frac{A}{n_{-1}}\right)^2 - 2B\left(\frac{A}{n_{-1}}\right) + H = 0,$$

which has the solutions,

$$\frac{A}{n_{-1}} = \frac{B \pm n_{-1}\left[\left(\frac{B}{n_{-1}}\right)^2 - cH \right]^{1/2}}{c\,n_{-1}^2}.$$

As $c \to 0$, that is as the spherical surface approaches a plane surface, $A \to 0$ as can be seen from Figure 5.6. To insure this we can use only the negative sign in the above solutions. A has the same sign as c; this can be seen either by considering the expression for A/n_{-1}, or from Figure 5.6. When A is negative, the tangent plane lies to the right of the surface. The coefficients B and H were introduced for convenience in calculation. Their physical significance is not difficult to understand. From the definition of H, and from Figure 5.7, it is seen that

$$H = c\,(X_T^2 + Y_T^2) = r\left[\frac{X_T^2 + Y_T^2}{r^2}\right] = r \tan^2 \beta,$$

where β is the angle between the optical axis and a line drawn from the center of curvature to the intersection of the ray with the tangent plane. From this expression for H, and the result derived in paragraph 5.4.4.4, an expression for B in terms of n_{-1}, and angles I and β can be found.

5.4.4.6 Before simplifying the expression for $\frac{A}{n_{-1}}$ a discussion of the physical meaning of the square root, $\left[(B/n_{-1})^2 - cH\right]^{1/2}$, is in order. This term will be used by itself in the refraction procedure; it is convenient to put it in another form here. Consider Figure 5.8; all the lines are in the plane of incidence. Using the cosine law, it can be stated that

$$D^2 = X_T^2 + Y_T^2 + r^2 = A^2 + r^2 + 2Ar\cos I.$$

Solving for $\cos I$, and substituting H for $c\,(X_T^2 + Y_T^2)$ produces

$$n_{-1}\cos I = \frac{H - c\,n_{-1}^2\left(\frac{A}{n_{-1}}\right)^2}{2\,\frac{A}{n_{-1}}}.$$

Finally, substituting the expression for $\frac{A}{n_{-1}}$, with the negative sign, given in paragraph 5.4.4.5, gives

$$n_{-1}\cos I = n_{-1}\left[\left(\frac{B}{n_{-1}}\right)^2 - cH\right]^{1/2}. \qquad (7)$$

5.4.4.7 Returning to the solution for $\frac{A}{n_{-1}}$ in paragraph 5.4.4.5, and using the expression for $n_{-1} \cos I$, we have

$$\frac{A}{n_{-1}} = \frac{B - n_{-1} \cos I}{c\, n_{-1}^2} .$$

But by using Equation (7)

$$c\, n_{-1}^2 = \frac{B^2 - n_{-1}^2 \cos^2 I}{H} = \frac{(B + n_{-1} \cos I)(B - n_{-1} \cos I)}{H} ,$$

and the final expression for $\frac{A}{n_{-1}}$ becomes,

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I} . \tag{8}$$

5.4.4.8 The four equations, then, which are used to calculate $\frac{A}{n_{-1}}$ are, in the order used,

$$H = c\,(X_T^2 + Y_T^2) , \tag{9}$$

$$B = M_{-1} - c\,(Y_T L_{-1} + X_T K_{-1}) , \tag{10}$$

$$n_{-1} \cos I = n_{-1} \left[\left(\frac{B}{n_{-1}}\right)^2 - c\,H \right]^{1/2} , \tag{7}$$

and

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I} . \tag{8}$$

Equations (4), (5), and (6) are then used to calculate X, Y, and Z.

5.4.5 Refraction procedure at the spherical surface.

5.4.5.1 Now that X, Y and Z have been calculated, these values together with initial data K_{-1}, L_{-1}, and M_{-1}, can be used to determine K, L, and M, which specify the direction of the ray after refraction. The basic equations which will be employed are Equations 2-(3) and 2-(4).

5.4.5.2 In Section 2 it was shown that Equation 2-(3) has the following meaning: if vectors are drawn (refer to Figure 2.3) from the intersection point, in the direction of the incident and refracted rays respectively, and these vectors have lengths equal to n_o and n_1, then the closing side of the triangle is parallel to the normal to the surface, and is of length Γ.

5.4.5.3 We now redraw this figure considering the surface as the jth surface. This is shown in Figure 5.9, which is drawn in the plane of incidence. Thus, the radius of curvature of the surface is also in this plane. The line of length Γ is parallel to r. The unit vector $\vec{M}_1$ is the quotient of the vector parallel to the normal divided by r. Hence

$$\vec{M}_1 = c\left[(0 - X)\vec{i} + (0 - Y)\vec{j} + (r - Z)\vec{k} \right]$$
$$= c\left[- X\vec{i} - Y\vec{j} + (r - Z)\vec{k} \right] ,$$

where $\vec{i}$, $\vec{j}$, $\vec{k}$ are unit vectors along the coordinate axes. Using Equation 2-(3),

$$\vec{S}_1 - \vec{S}_o = - c X \Gamma \vec{i} - c Y \Gamma \vec{j} + c\,(r - Z)\,\Gamma \vec{k} .$$

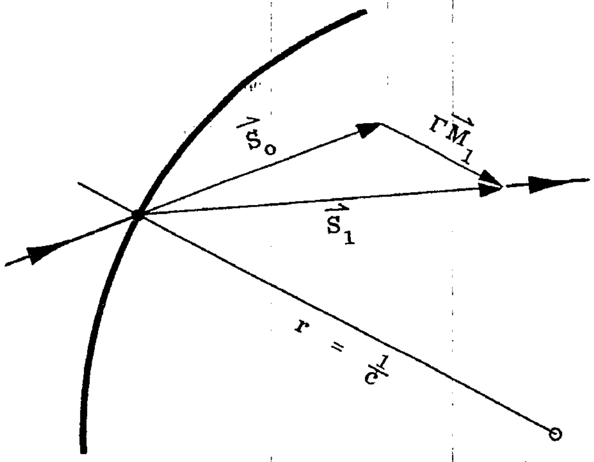


Figure 5.9 - Triangle for the law of refraction.

Now

$$\vec{S}_o = n_{-1}\vec{Q}_o = K_{-1}\vec{i} + L_{-1}\vec{j} + M_{-1}\vec{k},$$

and a similar equation holds for $\vec{S}_1$. Hence

$$\vec{S}_1 - \vec{S}_o = (K - K_{-1})\vec{i} + (L - L_{-1})\vec{j} + (M - M_{-1})\vec{k}.$$

Equating like coefficients of $\vec{i}$, $\vec{j}$, and $\vec{k}$, we have relations between the old and new optical direction cosines.

5.4.5.4 There remains the calculation of Γ. This is done by using Equation 2-(4). We can now write down the five equations which are used in the order given to calculate K, L, and M from the initial data or from previously calculated results.

$$n \cos I' = n\left[\left(\frac{n_{-1}}{n}\cos I\right)^2 - \left(\frac{n_{-1}}{n}\right)^2 + 1\right]^{1/2}, \qquad (11)$$

$$\Gamma = n \cos I' - n_{-1} \cos I, \qquad (12)$$

$$K = K_{-1} - Xc\Gamma, \qquad (13)$$

$$L = L_{-1} - Yc\Gamma, \qquad (14)$$

and

$$M = M_{-1} - (Zc - 1)\Gamma. \qquad (15)$$

5.4.6 Summary of ray trace equations.

5.4.6.1 In the previous sections there were derived the equations used to trace a skew ray from one surface through the following one. For convenience, the equations are now listed in the order of use. The initial ray data are X_{-1}, Y_{-1}, Z_{-1}, K_{-1}, L_{-1} and M_{-1}. The initial system data are t_{-1}, n_{-1} and c. Final values to be determined are X, Y, Z, K, L, and M.

$$\frac{d_{-1}}{n_{-1}} = (t_{-1} - Z_{-1}) \frac{1}{M_{-1}} , \tag{1}$$

$$Y_T = Y_{-1} + \frac{d_{-1}}{n_{-1}} L_{-1} , \tag{2}$$

$$X_T = X_{-1} + \frac{d_{-1}}{n_{-1}} K_{-1} , \tag{3}$$

$$H = c (X_T^2 + Y_T^2) , \tag{9}$$

$$B = M_{-1} - c (Y_T L_{-1} + X_T K_{-1}) , \tag{10}$$

$$n_{-1} \cos I = n_{-1} \left[\left(\frac{B}{n_{-1}} \right)^2 - c H \right]^{1/2} , \tag{7}$$

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I} , \tag{8}$$

$$X = X_T + \frac{A}{n_{-1}} K_{-1} , \tag{4}$$

$$Y = Y_T + \frac{A}{n_{-1}} L_{-1} , \tag{5}$$

$$Z = \frac{A}{n_{-1}} M_{-1} , \tag{6}$$

$$n \cos I' = n \left[\left(\frac{n_{-1}}{n} \cos I \right)^2 - \left(\frac{n_{-1}}{n} \right)^2 + 1 \right]^{1/2} , \tag{11}$$

$$\Gamma = n \cos I' - n_{-1} \cos I , \tag{12}$$

$$K = K_{-1} - X c \Gamma , \tag{13}$$

$$L = L_{-1} - Y c \Gamma , \tag{14}$$

and

$$M = M_{-1} - (Z c - 1) \Gamma . \tag{15}$$

5.4.6.2 The final calculated values, X, Y, Z, K, L, and M now become the initial ray data for the next calculation. The new system data, t, n, and c_{+1} must be given. These ray and system data are used with the above ray trace equations; in this way a given skew ray from any object surface can be traced through any number of spherical surfaces to the spherical image surface.

5.4.6.3 The equations listed in paragraph 5.4.6.1 are general, in that they also hold for plane surfaces. Referring to Figure 5.6, the physical result is that the j^th surface coincides with the tangent plane, hence the coordinates X_T, Y_T, Z_T equal X, Y, Z, and $A = 0$. These results follow mathematically by using $c = 0$ in the equations given in paragraph 5.4.6.1. Refraction at plane surfaces will be discussed in detail in Section 13.

5.4.7 Step by step ray tracing procedure.

5.4.7.1 The following table, Table 5.1, shows how these calculations can be made in a compact systematic manner. The surfaces are numbered 0, 1, 2, 3 beginning with 0 as the object surface. The initial system data are the values of the c, t, and n quantities indicated above the double line. In a numerical example (see Table 5.2) the values of these quantities are written in the places indicated. The letters in the left hand column have been defined in Section 5.2.2, or by the equations in Section 5.4.6.

5.4.7.2 The initial ray data are numerical values of X_o, Y_o, Z_o, K_o, L_o, and M_o, which would be written at the place indicated. Note that quantities pertaining to surfaces are written within the column for the corresponding surface; quantities pertaining to the space between surfaces are written in a break in the corresponding vertical line. The numbers running from 1 to 17 are the steps in the calculation in the order they are made. The steps, except (7) and (14), correspond to the 15 equations, listed in order of steps, in

Section 5.4.6.1. "Next step" indicates step No. 1 for the next ray segment. The table entries have been so chosen that a person using a desk calculator does not have to write down any number except those to be entered in the table.

SURFACE	0		1		2		3
c	c_o		c_1		c_2		
t		t_o		t_1		t_2	
n		n_o		n_1		n_2	
X	X_o		(9)				
Y	Y_o		(10)				
Z	Z_o		(11)				
K		K_o		(15)			
L		L_o		(16)			
M		M_o		(17)			
d_{-1}/n_{-1}		(1)		Next Step			
X_T			(2)				
Y_T			(3)				
H			(4)				
B			(5)				
$n_{-1}\cos I$			(6)				
$B + n_{-1}\cos I$			(7)				
A/n_{-1}			(8)				
$n \cos I'$			(12)				
Γ			(13)				
$c\,\Gamma$			(14)				

Table 5.1-Skew ray trace computing sheet.

SURFACE	0		1		2		3
c	0		0.25284872		-0.01473947		
t		-2.2		0.6			
n		1.0		1.62		1.0	
X	1.48		1.48		1.43679417		
Y	0		-0.33445977		-0.29386784		
Z	0		0.30264162		-0.01585220		
K		0		-0.24330257		-0.25700617	
L		0.17360000		0.22858306		0.23138586	
M		0.98481625		1.58522985		0.93830084	
d_{-1}/n_{-1}		-2.23391927		0.18758061			
X_T			1.48		1.43436116		
Y_T			-0.38780839		-0.29158202		
H			0.59186710		-0.03157802		
B			1.00183892		1.57910362		
$n_{-1}\cos I$			0.92413654		1.57871680		
$B + n_{-1}\cos I$			1.92597546		3.15782041		
A/n_{-1}			0.30730770		-0.00999994		
$n \cos I'$			1.57430250		0.93163659		
Γ			0.65016596		-0.64708020		
$c\,\Gamma$			0.16439363		0.00953762		

Table 5.2-Skew ray trace for three surfaces.

5.4.8 Numerical example.

5.4.8.1 Throughout the discussion of geometrical optics, lengthy explanations have been avoided by the inclusion of numerical examples showing the actual calculations. Table 5.2 is such an illustration. The calculations shown in this table can be made by an experienced person with a modern desk calculator without undue labor. In order for the calculations to be useful, at least six significant figures must be carried throughout. Since the introduction of the modern electronic computing machines, there is really very little justification for a human computer to carry out these calculations unless ray tracing is only done occasionally. The above equations can be programmed in a modern machine to make these calculations in less than one second per surface, with at least eight significant figures. The calculations shown in this and other numerical examples may not offer complete consistency in the number of significant figures for two reasons: (1) some were prepared from automatic computer results where intermediate values were not available and had to be developed by hand computing; (2) others were prepared from a designer's work sheets where the aim was not eight-figure accuracy but only three or four-figure accuracy in which case the designer had merely entered results as they appeared on the hand calculator. No units appear in this and other numerical examples, because the equations are valid for any set of consistent units. As long as all lengths are in the same units, the numerical example will be correct for any units.

5.4.8.2 Some specific remarks should be made about Table 5.2 . The initial data which are given to one, two, or three significant figures are assumed exact. From the initial system data it is apparent that we are considering a double convex lens, index 1.62, surrounded by air. The incident light first intersects the convex face of the lens. The lens thickness is about one quarter of the distance between lens and object surface, but no information is given (or needed for a ray trace) concerning the absolute magnitude of any distance. The object surface is to the right of the lens; therefore the object is virtual.

5.4.8.3 From the initial ray data we see that the (virtual) object point, that is the point towards which the ray is heading, is on the X axis, but not in the Y - Z plane. The initial ray is parallel to the Y - Z plane, hence $X_1 = X_0$. The ray is inclined upwards at an angle with the Z axis of 10°. The calculations indicate that the ray intersects both surfaces of the lens below the X - Z plane (because Y is negative), and intersects both surfaces at points "away from the reader" with respect to the Y - Z plane (because X is positive). The Z value at the first surface is positive because the curvature is positive ; likewise the Z value at the second surface is negative.

5.5 SKEW RAY TRACE EQUATIONS FOR ASPHERIC SURFACES

5.5.1 General.

5.5.1.1 The discussion in Section 5.4 developed equations for, and demonstrated their use in, ray tracing procedures through spherical surfaces. Although spherical surfaces are still much easier to make, and hence are preferred by the lens maker, aspheric surfaces are readily handled by the lens designer who has access to an electronic computer. Aspheric surfaces afford the designer a great deal more latitude in the design, and in addition often permit better correction of aberrations. Aspheric surfaces are being used more and more, and their widespread use depends on inexpensive methods of production.

5.5.1.2 In the skew ray trace for spherical surfaces, it was convenient to effect the transfer from one physical surface to the next by introducing a non-physical tangent plane, and effecting the transfer in two steps. In the case of aspheric surfaces we introduce two non-physical surfaces, a plane and a sphere, both tangent to the physical aspheric surface at the optical axis. See Figure 5.10. The transfer between physical surfaces is now effected in three steps:

(1) first surface to next tangent plane;

(2) tangent plane to tangent sphere;

(3) tangent sphere to physical (second) surface.

Steps (1) and (2) are carried out using the procedure already developed in Section 5.4 .

5.5.2 Mathematical description of an aspheric surface.

5.5.2.1 We need to describe the aspheric surface in a way that indicates clearly its departure from the tangent sphere. This kind of description will not only be easily handled by the ray trace equations, but will also quickly and quantitatively show how close in form the aspheric is to the sphere.

5.5.2.2 In Paragraph 5.4.4.4 there is given an equation for Z ; this quantity is called the sag of the sphere, an abbreviation of sagitta. Using $S^2 = X^2 + Y^2$, this equation is

$$Z = \frac{1}{c} \left[1 - (1 - c^2 S^2)^{1/2} \right] .$$

$c = \frac{1}{R}$

By multiplying and dividing by $\left[1 + (1 - c^2 S^2)^{1/2} \right]$, we have

$$Z = \frac{c\, S^2}{1 + \sqrt{1 - c^2 S^2}} .$$

Because the shape of an aspheric surface (which is assumed to have rotational symmetry about the Z axis) differs from that of the tangent sphere, the sag (Z) of the aspheric at any distance S from the axis may differ from the sag of the tangent sphere. This is indicated by expressing the difference in these two sags by a power series in S^2 . (The series is in powers of S^2 , and hence only even powers of S appear, because the aspheric has rotational symmetry about the Z axis.) The final expression for the sag is

$$Z = \frac{c S^2}{1 + \sqrt{1 - c^2 S^2}} + eS^4 + fS^6 + gS^8 + hS^{10} + O(S^{12})$$

5.5.2.3 Each of the numerical coefficients e , f , g , and h may be positive or negative. The term $O(S^{12})$ stands for the rest of the series, that is terms of order 12 and higher. In a numerical calculation, if the sag is given by this expression, $O(S^{12})$ would be assumed zero, and the calculations would involve only e , f , g , and h . The terms eS^4 , fS^6 , etc., are called deformation terms.

5.5.3 Initial data, and transfer from physical surface to next tangent sphere. Part of the transfer from one physical surface to the next has already been solved in Section 5.4 . The initial ray data for the skew ray between aspheric surfaces is the same as given in Section 5.4.2 , namely X_{-1} , Y_{-1} , Z_{-1} , K_{-1} , L_{-1} and M_{-1} . We determine the intersection of this ray with the non-physical sphere, tangent to the jth aspheric surface, by the procedure given in Sections 5.4.3 and 5.4.4 . In other words we apply Equations (1), (2), (3), (9), (10), (7), (8), (4), (5) and (6) in that order. The only difference so far between this ray trace and the former is that in the previous case the sphere was a physical surface, while in the present case it is a purely fictitious surface. The equations do not know the difference between physical and non-physical surfaces; hence the same equations are used for both cases.

5.5.4 Transfer procedure, tangent sphere to aspheric surface.

5.5.4.1 In Paragraphs 5.4.4.4 and 5.4.4.5 an expression for $\frac{A}{n_{-1}}$ was derived using four equations, namely the equation for the sag, Z , of the sphere and Equations (4), (5), and (6) . This value of $\frac{A}{n_{-1}}$ was then used in Equations (4), (5), and (6) to transfer from tangent plane to sphere. It would be perfectly possible to proceed similarly here. We would set up three equations, corresponding to (4), (5), and (6) , but replacing A by A + A' . (See Figure 5.10) . Using these three equations, and the equation for the

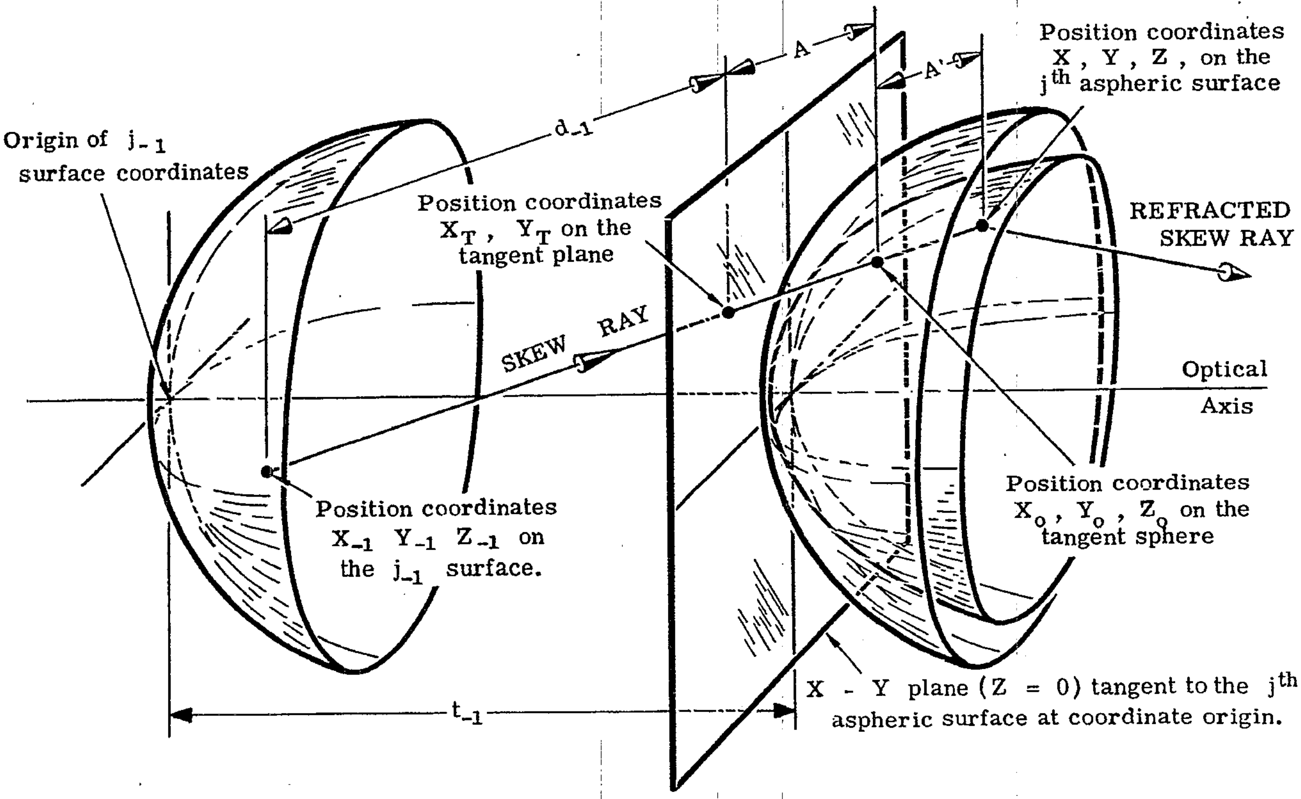


Figure 5.10 - Diagram of a skew ray in space between the j_{-1} surface and the j^{th} aspheric surface.

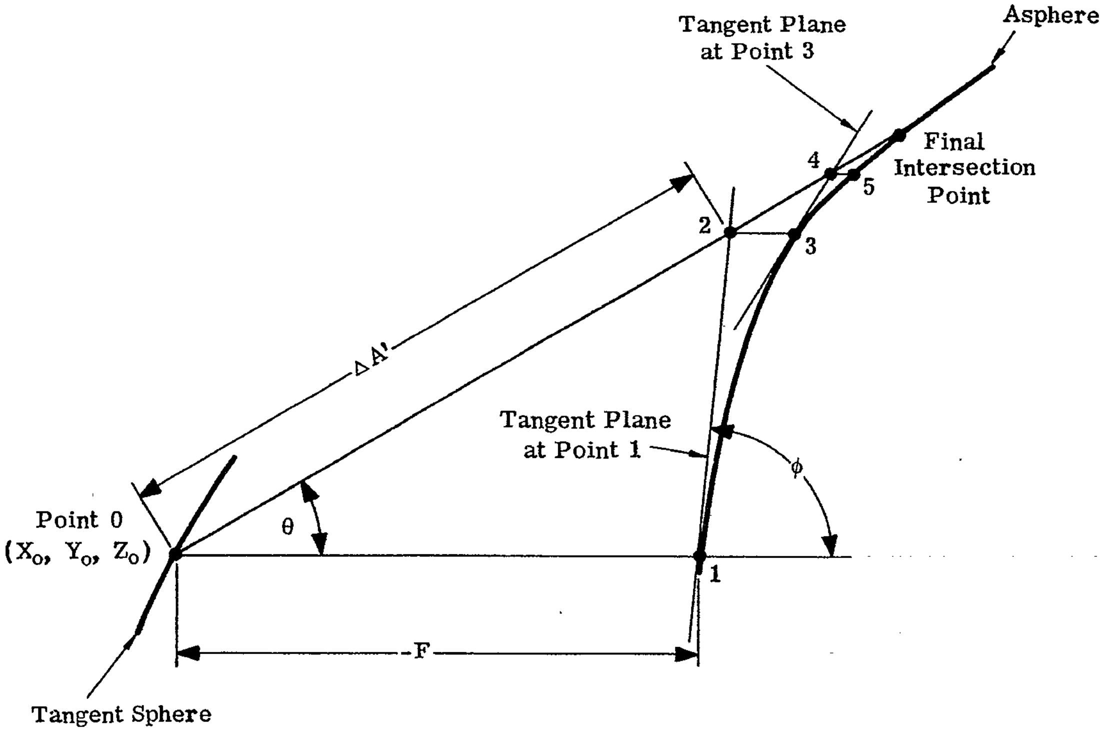


Figure 5.11- Step-wise approximations from tangent sphere intersection, point 0, to final intersection point.

sag of an aspheric surface, Paragraph 5.5.2.2, an expression for $\frac{A + A'}{n_{-1}}$ could be found, and used to transfer directly from tangent sphere to aspheric. The resulting calculations are extremely involved and it is preferable to proceed otherwise.

5.5.4.2 The procedure to be employed makes use of the fact that transfer to the tangent sphere is fairly simple. The remaining transfer from tangent sphere to aspheric is effected in a step-wise procedure approaching the final intersection by successive approximations. The physical procedure is indicated in Figure 5.11 ; this figure represents the plane determined by the skew ray, and a line through this ray parallel to the Z axis.

5.5.4.3 Beginning at point 0 , the intersection of the ray with the tangent sphere, the first approximation to the final point is point 1 . Point 1 has the same X and Y values as point 0 ; its Z value differs from that of point 0 by the deformation terms evaluated at these particular values of X and Y . The second approximation is point 2 , the intersection of the ray with a line tangent to the aspheric at 1 . The tangent line is determined from the known coordinates of point 1 and the calculated curvature of the aspheric at point 1 . Since the ray direction is known, its intersection with the tangent line, point 2 , is determined. The procedure is now repeated. Point 3 has the same X and Y as point 2 , and its Z value can be found from the deformation terms and the Z value at point 2 . The fourth and fifth approximations are points 4 and 5 , respectively. (The point 0 , on the sphere, is correctly called the zeroth approximation to the final point.)

5.5.4.4 The various values of X , Y , and Z for points 0, 1, 2,··· will be referred to as X_n , Y_n , and Z_n where the n will stand for the order of approximation. Let us begin at any even-numbered point, that is a point on the ray; in practice, the calculations begin with point 0 , but we wish to make the equations general so that n will stand for any even-numbered point. The next point, on the aspheric, will have coordinates X_n , Y_n , Z_m . Note that the X and Y values are the same as for the previous point. The S_n value now used to calculate the sag, Z_m , is

$$S_n^2 = X_n^2 + Y_n^2 . \qquad (16)$$

The change in Z , that is $Z_m - Z_n$, is the distance parallel to the Z axis between an even-numbered point and an odd-numbered point. Calling this distance $-F$ (see Figure 5.11), we have

$$F = Z_n - \left[\frac{c\,S^2}{1 + \sqrt{1 - c^2 S^2}} + eS^4 + fS^6 + gS^8 + hS^{10} \right]. \qquad (17)$$

(The subscript n has been omitted. Henceforth all values of S are rigorously S_n.)

5.5.4.5 For notational purposes it is convenient to designate the square root in Equation (17) by W. Hence

$$W = \left[1 - c^2 S^2 \right]^{1/2}. \qquad (18)$$

Referring to Figure 5.7, it can be seen that $W = \sqrt{1 - \sin^2 \alpha} = \cos \alpha$, where α is the angle between the normal to the surface and the optical axis. W therefore is the direction cosine, with respect to the Z axis, of a radius drawn from the sphere to the center.

5.5.4.6 From an odd-numbered point, whose coordinates we now know, we move along the tangent line to the ray. The coordinates of this new even-numbered point are X_{n+1}, Y_{n+1}, and Z_{n+1}. Calling the distance along the ray, between two even-numbered points, $\Delta A'$, we can write equations for the new coordinates similar to equations (4), (5), and (6). We have then

$$X_{n+1} = X_n + \frac{\Delta A'}{n_{-1}} K_{-1}, \qquad (19)$$

$$Y_{n+1} = Y_n + \frac{\Delta A'}{n_{-1}} L_{-1}, \qquad (20)$$

and

$$Z_{n+1} = Z_n + \frac{\Delta A'}{n_{-1}} M_{-1}. \qquad (21)$$

We will consider the calculation of $\Delta A'$ presently. Once this is known, the new coordinates on the ray are known, and we repeat the calculations through two more steps until we get once again to the ray. This iteration procedure is continued until $\Delta A'/n_{-1}$ is less than any desired tolerance. In this manner we can approach the final point on the aspheric as closely as we choose.

5.5.4.7 The remaining problem in the transfer from tangent sphere to aspheric surface is the determination of $\Delta A'$. First we need the equation of the plane tangent to the aspheric surface at an odd-numbered point. From the equation for the sag of the surface, Paragraph 5.5.2.2, we can write

$$\psi(X, Y, Z) = Z - \left[\frac{c\,S^2}{1 + \sqrt{1 - c^2 S^2}} + eS^4 + fS^6 + gS^8 + hS^{10} \right] = 0,$$

where $\psi(X, Y, Z) = 0$ is the equation of the aspheric surface. Now a plane, tangent to the surface at the point X_n, Y_n, Z_m will coincide with the first approximation to the surface. Physically, if we restrict ourselves to points close to (X_n, Y_n, Z_m) the surface is a plane. To find the first approximation to the surface we expand $\psi(X, Y, Z)$ and keep only the zeroth and first order terms.

5.5.4.8 The equation of the tangent plane is then

$$\psi(X_n, Y_n, Z_m) + (X - X_n)\left[\frac{\partial \psi}{\partial X}\right]_{X_n, Y_n, Z_m} + (Y - Y_n)\left[\frac{\partial \psi}{\partial Y}\right]_{X_n, Y_n, Z_m} + (Z - Z_m)\left[\frac{\partial \psi}{\partial Z}\right]_{X_n, Y_n, Z_m} = 0,$$

where the first term is the zeroth order term, and the last three are the first order terms, in the expansion of $\psi(X, Y, Z)$. Using Equation (16) we have

$$\frac{\partial \psi}{\partial X} = \frac{\partial \psi}{\partial S}\frac{\partial S}{\partial X} = \frac{\partial \psi}{\partial S}\frac{X}{S},$$

$$\frac{\partial \psi}{\partial Y} = \frac{\partial \psi}{\partial S}\frac{Y}{S},$$

and

$$\frac{\partial \psi}{\partial Z} = 1.$$

Using the expression for $\psi(X, Y, Z)$ given in Paragraph 5.5.4.7, and Equation (18), we get

$$\frac{\partial \psi}{\partial S} = -\frac{S}{W} c - S \left[4eS^2 + 6fS^4 + 8gS^6 + 10hS^8 \right],$$

which can be simplified to $\frac{\partial \psi}{\partial S} = -\frac{S}{W} E$ by defining

$$E = c + W \left[4eS^2 + 6fS^4 + 8gS^6 + 10hS^8 \right]. \qquad (22)$$

(If the deformation coefficients are small, E is approximately the curvature of the aspheric surface at the distance S_n from the optical axis.)

5.5.4.9 The equation for the sag of the aspheric surface given in Paragraph 5.5.2.2 is an equation for Z_m if $S^2 = X_n^2 + Y_n^2$. Because of this, $\psi(X_n, Y_n, Z_m)$ is zero, using the equation for ψ in Paragraph 5.5.4.7. The zeroth order term in the expansion is therefore zero. The equation of the plane becomes, using the above expressions for the partial derivatives,

$$-(X - X_n)\frac{X_n}{W} E - (Y - Y_n)\frac{Y_n}{W} E + (Z - Z_n)\frac{W}{W} + Z_n - Z_m = 0,$$

where we have separated the term $Z - Z_m$ into two terms. By Equation (17), $F = Z_n - Z_m$. We define here two quantities,

$$U = -XE, \qquad (23)$$

$$V = -YE. \qquad (24)$$

5.5.4.10 With these substitutions the equation of the plane becomes

$$(X - X_n)U + (Y - Y_n)V + (Z - Z_n)W = -FW.$$

This equation holds for all values of X, Y, and Z, in particular X_{n+1}, Y_{n+1}, and Z_{n+1}. Instead of the difference $(X_{n+1} - X_n)$, we use $(\Delta A'/n_{-1})K_{-1}$ from Equation (19). Similarly, using Equations (20) and (21), and solving for $\Delta A'/n_{-1}$,

$$\frac{\Delta A'}{n_{-1}} = \frac{-FW}{K_{-1}U + L_{-1}V + M_{-1}W}. \qquad (25)$$

From Figure 5.10, it can be seen that the distance, D_{-1}, along the ray is

$$D_{-1} = d_{-1} + A + A'. \qquad (25a)$$

5.5.5 Refraction procedure at the aspheric surface.

5.5.5.1 Now that the intersection point, (X, Y, Z), of the ray and the aspheric surface has been found, the refraction equation is used to determine the new direction of the ray. The procedure is basically the same as that used for refraction at spherical surfaces, discussed in Section 5.4.5. In that section Equation (11) was used to calculate $n \cos I'$, because $n_{-1} \cos I$ had already been calculated using Equation (7).

5.5.5.2 In the present case there is not yet a value for $n_{-1} \cos I$. To calculate this we use the fact that the cosine of an angle between two directed lines is equal to the sum of the products of their corresponding direction cosines. Since we are calculating $\cos I$, the two lines in question are the ray whose optical direction cosines are K_{-1}, L_{-1}, and M_{-1}, and the normal to the aspheric surface. Now the normal to the surface is just the normal to the tangent plane. The equation of this tangent plane is given in Paragraph 5.5.4.10, where X_n, Y_n, and Z_n are the coordinates of the final point on the ray, the intersection with the surface.

5.5.5.3 Given the equation of a plane, the direction cosines of the normal are proportional to the corresponding coefficients of X, Y, and Z. Hence the direction cosines of the normal are, in the usual order, U/G, V/G, and W/G, where G is a proportionality constant. Because the sum of the squares of the direction cosines is unity, we have

$$G^2 = U^2 + V^2 + W^2. \qquad (26)$$

Using the direction cosines of the ray we get

$$\cos I = \frac{K_{-1}}{n_{-1}} \frac{U}{G} + \frac{L_{-1}}{n_{-1}} \frac{V}{G} + \frac{M_{-1}}{n_{-1}} \frac{W}{G} ,$$

which is rewritten in final form as

$$G\, n_{-1} \cos I = K_{-1}\, U + L_{-1}\, V + M_{-1}\, W . \qquad (27)$$

5.5.5.4 Equation (11) is now used to determine $n \cos I'$. However, for calculation purposes, it is preferable to leave the G on both sides of the equation.

$$G\, n \cos I' = n \left[\left(G \frac{n_{-1}}{n} \cos I \right)^2 - G^2 \left(\frac{n_{-1}}{n} \right)^2 + G^2 \right]^{1/2} . \qquad (28)$$

5.5.5.5 Returning to the equation in Paragraph 5.4.5.3, we write this vector equation as three scalar equations using the method of Paragraph 5.4.5.4. We get

$$K - K_{-1} = \Gamma \frac{U}{G} ,$$

$$L - L_{-1} = \Gamma \frac{V}{G} ,$$

and

$$M - M_{-1} = \Gamma \frac{W}{G} ,$$

because Γ is parallel to the normal to the surface and therefore has the same direction cosines. Introducing $P = \Gamma/G$, we have, using Equation (12),

$$P = (G\, n \cos I' - G\, n_{-1} \cos I)/G^2 . \qquad (29)$$

Finally, K, L, and M are found from the equations,

$$K = K_{-1} + U P , \qquad (30)$$

$$L = L_{-1} + V P , \qquad (31)$$

and

$$M = M_{-1} + W P . \qquad (32)$$

5.5.6 Summary of ray trace equations.

5.5.6.1 In the previous sections we have derived the equations used to trace a skew ray from a tangent sphere through the aspheric surface. For convenience we rewrite the equations in the order of use. The initial ray data are X_{-1}, Y_{-1}, Z_{-1}, K_{-1}, L_{-1}, and M_{-1}. The initial system data are t_{-1}, n_{-1}, c, and the deformation coefficients e, f, g, $\cdots$. Final values to be determined are X, Y, Z, K, L, and M.

5.5.6.2 The position coordinates for the ray on the tangent sphere are calculated using the first ten equations listed in Section 5.4.6. Equations (16) through (32) are then used in the order listed below.

$$S_n^2 = X_n^2 + Y_n^2 , \qquad (16)$$

$$W = \left[1 - c^2 S^2 \right]^{1/2} , \qquad (18)$$

$$F = Z_n - \left[\frac{c\, S^2}{1 + \sqrt{1 - c^2 S^2}} + eS^4 + fS^6 + gS^8 + hS^{10} \right] , \qquad (17)$$

$$E = c + W \left[4eS^2 + 6fS^4 + 8gS^6 + 10hS^8 \right] , \qquad (22)$$

$$U = -XE , \qquad (23)$$

$$V = -YE , \qquad (24)$$

$$\frac{\Delta A'}{n_{-1}} = \frac{-FW}{K_{-1}\ U + L_{-1}\ V + M_{-1}\ W}\ , \tag{25}$$

$$X_{n+1} = X_n + \frac{\Delta A'}{n_{-1}}\ K_{-1}\ , \tag{19}$$

$$Y_{n+1} = Y_n + \frac{\Delta A'}{n_{-1}}\ L_{-1}\ , \tag{20}$$

$$Z_{n+1} = Z_n + \frac{\Delta A'}{n_{-1}}\ M_{-1}\ , \tag{21}$$

$$G^2 = U^2 + V^2 + W^2\ , \tag{26}$$

$$G\ n_{-1}\ \cos I = K_{-1}\ U + L_{-1}\ V + M_{-1}\ W\ , \tag{27}$$

$$G\ n\ \cos I' = n\left[\left(G\ \frac{n_{-1}}{n}\cos I\right)^2 - G^2\left(\frac{n_{-1}}{n}\right)^2 + G^2\right]^{1/2}\ , \tag{28}$$

$$P = (G\ n\ \cos I' - G\ n_{-1}\ \cos I)/G^2\ , \tag{29}$$

$$K = K_{-1} + UP\ , \tag{30}$$

$$L = L_{-1} + VP\ , \tag{31}$$

and

$$M = M_{-1} + WP\ . \tag{32}$$

5.5.6.3 The first ten of these equations are used in an iterative process until $\Delta A'/n_{-1}$ becomes as small as desired. The final values of U, V, and W are then used in the last seven equations (26) through (32). The final calculated values of X, Y, Z, K, L, and M become the initial ray data for the next calculation. These values, together with new system data, t, n, c_{+1}, and deformation terms, are used in a reapplication of the ray trace equations.

SURFACE	0	1	2	3
c	0	0.25284872	-0.01473947	
e		-0.005		
f		0.00001		
g		-0.0000005		
h		0		
t		-2.2	0.6	
n		1.0	1.62	1.0
X	1.48	1.48	1.44043943	
Y	0	-0.33905030	-0.29645624	
Z	0	0.27660001	-0.01594078	
K		0	-0.20481560	
L		0.17360000	0.22052072	
M		0.98481625	1.59179807	
d_{-1}/n_{-1}		-2.23391927		
X_T		1.48		
Y_T		-0.38780839		
H		0.59186710		
B		1.00183892		
$n_{-1} \cos I$		0.92413654		
$B + n_{-1} \cos I$		1.92597546		
A/n_{-1}		0.30730770		
$n \cos \Gamma$				
Γ		Enter $X_n\ Y_n\ Z_n$		
$c\ \Gamma$		in Table 5.4		

Table 5.3 - Skew ray trace through an aspheric surface. Part of the calculations are shown in Table 5.4.

5.5.6.4 Because a spherical surface is a special case of an aspheric surface for which the deformation terms are zero, the ray trace equations for aspheric surfaces should easily reduce to those for spherical surfaces. We see, for the case of a sphere $(e = f = g = h = \dots = 0)$,

$$E = c ,$$

$$U = - X c ,$$

$$V = - Y c ,$$

$$W = - (Z c - 1) , \quad \text{(holds for aspheric also)}$$

$$G = 1 ,$$

$$n_{-1} \cos I = - c \left[X K_{-1} + Y L_{-1} + Z M_{-1} \right] ,$$

$$P = \Gamma ,$$

and equations (30), (31), and (32) become identical with equations (13), (14), and (15).

ITERATION	1	2	3
X_n	1.48000000	1.48000000	1.48000000
Y_n	-0.33445977	-0.33905071	-0.33905030
Z_n	0.30264163	0.27659764	0.27659999
S_n^2	2.30226334	2.30535538	2.30535510
$1 - c^2 S^2$	0.85281060	0.85261290	0.85261290
W	0.92347745	0.92337040	0.92337040
$c / (1 + W)$	0.13145395	0.13146127	0.13146127
hS^2	0.00000000	0.00000000	0.00000000
$hS^4 + gS^2$	-0.00000115	-0.00000115	-0.00000115
$hS^6 + gS^4 + fS^2$	0.00002037	0.00002040	0.00002040
$hS^8 + gS^6 + fS^4 + eS^2$	-0.01146441	-0.01147976	-0.01147976
$hS^{10} + gS^8 + fS^6 + eS^4 + \frac{cS^2}{1 + W}$	0.27624751	0.27660004	0.27660000
$-F$	-0.02639412	-0.00000238	-0.00000002
$-10\ hS^2$	0.00000000	0.00000000	0.00000000
$-10\ hS^4 - 8\ gS^2$	0.00000921	0.00000922	0.00000922
$-10\ hS^6 - 8\ gS^4 - 6\ fS^2$	-0.00011693	-0.00011706	-0.00011706
$10\ hS^8 + 8\ gS^6 + 6\ fS^4 + 4\ eS^2$	-0.04577605	-0.04583724	-0.04583723
$-E$	-0.21057557	-0.21052397	-0.21052398
U	-0.31165184	-0.31157548	-0.31157549
V	0.07042906	0.07137830	0.07137822
$K_{-1}U + L_{-1}V + M_{-1}W$	0.92168208	0.92174144	0.92174143
$-FW$	-0.02437438	0.000002198	0.000000018
$\Delta A'/n_{-1}$	-0.02644553	0.000002384	0.000000020
X	1.48000000	1.48000000	1.48000000
Y	-0.33905071	-0.33905030	-0.33905030
Z	0.27659764	0.27659999	0.27660001
G^2			0.95478704
$Gn \cos I'$			1.54937517
P			0.65735466
K			-0.20481560
L			0.22052072
M			1.59179807

Table 5.4 - Skew ray trace iteration and refraction calculations. The table shows three iterations.

5.5.7 Numerical example.

5.5.7.1 A numerical example is shown in Tables 5.3 and 5.4. The system data is the same as the example shown in Table 5.2, except for the addition of three deformation coefficients e, f, and g. The coefficient h is specifically listed in both tables as zero. This avoids possible error in not being certain whether or not a coefficient was erroneously omitted. The initial ray data is identical with the previous example; hence the calculations and results for transfer to the tangent sphere are the same. Thus steps 1 through 11 (see Table 5.1) are identical, except for the location of the results of step 11. These are placed in Table 5.4 and are the initial data for the iteration process.

5.5.7.2 Table 5.4 shows the iteration process by which ($\Delta A'/n_{-1}$) < 0.00001 ; this represents the criterion, set up prior to the calculations, to determine when the iteration process is to be stopped. It is noticed that the first value of $\Delta A'/n_{-1}$ is negative, the second positive, the third almost zero. This oscillation about the target value (< 0.00001) is typical of the method of successive approximations. This method will be used in later sections where aberrations are discussed.

5.5.7.3 The final values of X , Y , and Z , shown just above the double line in Table 5.4 in the column 3 , are entered in Table 5.3 in the place for steps 9 , 10 , and 11 . (The entire iteration process gives the results for steps 9 , 10 , and 11 for an aspheric surface.) These values are now part of the initial ray data for the next surface. The refraction calculations at the aspheric surface are given in Table 5.4 below the double line, and use the final results found above. The values of K , L , and M are now entered in Table 5.3 as the results of steps 15 , 16 , and 17 . They will be used as initial data for the next surface.

5.6 MERIDIONAL RAYS

5.6.1 Definition. A meridional ray is any ray lying in a plane containing the optical axis. A meridional ray will remain in the same plane throughout an entire centered system. For this reason, the tracing of meridional rays is a two dimensional problem, while the tracing of skew rays, which do not lie in a plane containing the optical axis, is a three dimensional problem.

5.6.2 Use of skew ray trace equations. The skew ray formulae given in Sections 5.4 and 5.5 are designed for use on modern automatic computing machines. However, they are in a form which can be used with relative ease - for skew rays - on a desk calculator. Extensive skew ray tracing, which is essential in order to make a complete analysis of a lens system, should be done on a computing machine. In the preliminary design of a lens system it is usually convenient to trace a few selected meridional rays. These are often traced by hand. If the object point has coordinates ($X_o = 0$, Y_o , Z_o) and the ray pierces the first surface at coordinates ($X_1 = 0$, Y_1 , Z_1) the ray is meridional and will remain in the YZ-plane all the way to the image surface. Meridional rays can be traced using the skew ray formulae given in Sections 5.4 and 5.5 by setting $X = 0$ and $K = 0$.

5.6.3 Meridional ray trace, spherical surfaces.

5.6.3.1 Meridional ray tracing can be done for spherical surfaces by using Equations (1) through (10) , followed by either Equations (11) through (15) or Equations (16) through (32). For meridional rays, Equations (1) through (10) reduce to the following eight equations, in the order used:

$$\frac{d_{-1}}{n_{-1}} = (t_{-1} - Z_{-1}) \frac{1}{M_{-1}} , \tag{1}$$

$$Y_T = Y_{-1} + \frac{d_{-1}}{n_{-1}} L_{-1} , \tag{2}$$

$$H = c Y_T^2 , \tag{9a}$$

$$B = M_{-1} - c Y_T L_{-1} , \tag{10a}$$

$$n_{-1} \cos I = n_{-1}\left[\left(\frac{B}{n_{-1}}\right)^2 - cH\right]^{1/2} , \tag{7}$$

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I} , \tag{8}$$

$$Y = Y_T + \frac{A}{n_{-1}} L_{-1} , \tag{5}$$

and

$$Z = \frac{A}{n_{-1}} M_{-1} . \tag{6}$$

Only eight equations are needed, the other two being $X_T = X = 0$. These eight equations trace a meridional ray from any surface to the next spherical surface.

5.6.3.2 Refraction at the spherical surface may be calculated by applying Equations (11), (12), (14), and (15) as written. Equation (13) becomes $K = 0$. (This procedure is referred to as the short form). On the other hand Equations (16) through (32) may be used. These are reduced to the following seven equations, in the order used:

$$W = \left[1 - c^2 Y^2 \right]^{1/2} , \tag{18a}$$

$$V = - Y c , \tag{24a}$$

$$n_{-1} \cos I = L_{-1} V + M_{-1} W , \tag{27a}$$

$$n \cos I' = n \left[\left(\frac{n_{-1}}{n} \cos I \right)^2 - \left(\frac{n_{-1}}{n} \right)^2 + 1 \right]^{1/2} , \tag{11}$$

$$\Gamma = n \cos I' - n_{-1} \cos I , \tag{12}$$

$$L = L_{-1} - Y c \Gamma , \tag{14}$$

and

$$M = M_{-1} - W \Gamma . \tag{32a}$$

Only seven equations are needed, the other ten being $S_n = Y_n$, $E = c$, $G = 1$, $Y_{n+1} = Y_n$, $Z_{n+1} = Z_n$, and $F = U = \Delta A' = X_{n+1} = K = 0$.

5.6.4 Meridional ray trace, aspheric surfaces. For meridional rays and aspheric surfaces, after applying the eight equations given in Paragraph 5.6.3.1, the Equations (16) through (32) are used. These reduce to the following thirteen equations, in the order used:

$$W = \left[1 - c^2 Y^2 \right]^{1/2} , \tag{18a}$$

$$F = Z_n - \left[\frac{c Y^2}{1 + W} + e Y^4 + f Y^6 + g Y^8 + h Y^{10} \right] , \tag{17a}$$

$$E = c + W \left[4 e Y^2 + 6 f Y^4 + 8 g Y^6 + 10 h Y^8 \right] , \tag{22a}$$

$$V = - Y E , \tag{24}$$

$$\frac{\Delta A'}{n_{-1}} = \frac{- F W}{L_{-1} V + M_{-1} W} , \tag{25b}$$

$$Y_{n+1} = Y_n + \frac{\Delta A'}{n_{-1}} L_{-1} , \tag{20}$$

$$Z_{n+1} = Z_n + \frac{\Delta A'}{n_{-1}} M_{-1} , \tag{21}$$

$$G^2 = V^2 + W^2 , \tag{26a}$$

$$G n_{-1} \cos I = L_{-1} V + M_{-1} W , \tag{27a}$$

$$G n \cos I' = n \left[\left(G \frac{n_{-1}}{n} \cos I \right)^2 - G^2 \left(\frac{n_{-1}}{n} \right)^2 + G^2 \right]^{1/2} , \tag{28}$$

$$P = (G n \cos I' - G n_{-1} \cos I) / G^2 , \tag{29}$$

$$L = L_{-1} + V P , \tag{31}$$

and

$$M = M_{-1} + W P . \tag{32}$$

Only 13 equations are needed, the other four being $S_n = Y_n$, and $U = X_{n+1} = K = 0$.

5.6.5 Simplified meridional ray trace, spherical surfaces.

5.6.5.1 There are many other methods, involving different parameters, which are commonly used to trace meridional rays. One such method specifies the angle the ray makes with the optical axis, and the perpendicular distance from the center of curvature of the surface to the ray. Figure 5.12 indicates the two

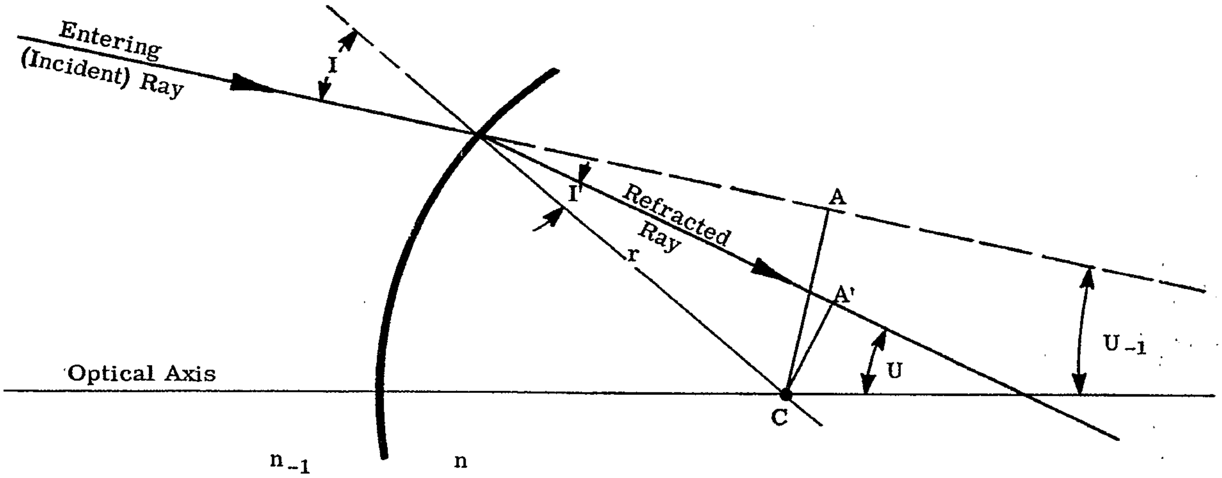


Figure 5. 12—Ray tracing by the PR method.

quantities specified, U_{-1} and CA. The corresponding quantities, U and CA', specify the refracted ray. This diagram involves the angles U and U_{-1}, called the slope angles. We will use a convention for the sign of a slope angle similar to that for incidence, reflection, and refraction angles. (See Section 2.2.2). If the ray must be rotated clockwise through the acute angle to bring it into coincidence with the optical axis the angle is called positive. Both U and U_{-1} are negative as drawn.

5.6.5.2 The following equations are readily derived from the figure: *

$$\sin I = \frac{CA}{r},$$

and

$$\sin I' = \frac{CA'}{r}.$$

Therefore from Snell's law

$$n_{-1} \sin I = \frac{CAn_{-1}}{r} = \frac{CA'n}{r} = n \sin I'.$$

By definition:

$$P = CAn_{-1} = CA'n,$$

and

$$R = \frac{1}{n_{-1}\, r}, \quad R' = \frac{1}{nr}.$$

(Because of these two definitions, this method is referred to as the PR method.)

---

* The notation used in this simplified ray trace must not be confused with the skew ray formulae. There has been no attempt to avoid duplication of symbols.

The refraction equations then become

$$\sin I = PR, \tag{33}$$

$$\sin I' = PR', \tag{34}$$

and

$$U = U_{-1} - (I - I'). \tag{35}$$

The value of P is transferred from one surface to the next by the following equation:

$$P_{+1} = P - (r - r_{+1} - t)\, n \sin U. \tag{36}$$

5.6.5.3 Equation (36) is seen to follow from Figure 5.13. We have

$$P_{+1} = C_{+1} A_{+1}\, n = C A'\, n + C C_{+1}\, n \sin U,$$

because U is negative. The distance $C C_{+1} = t - r + r_{+1}$, and Equation (36) follows by rearrangement. The above ray tracing equations, (33) through (36), require a minimum of calculation and are ideal for hand computing. If several rays are to be calculated it is worth while to precalculate the lens constants R, R', and $n(r - r_{+1} - t)$.

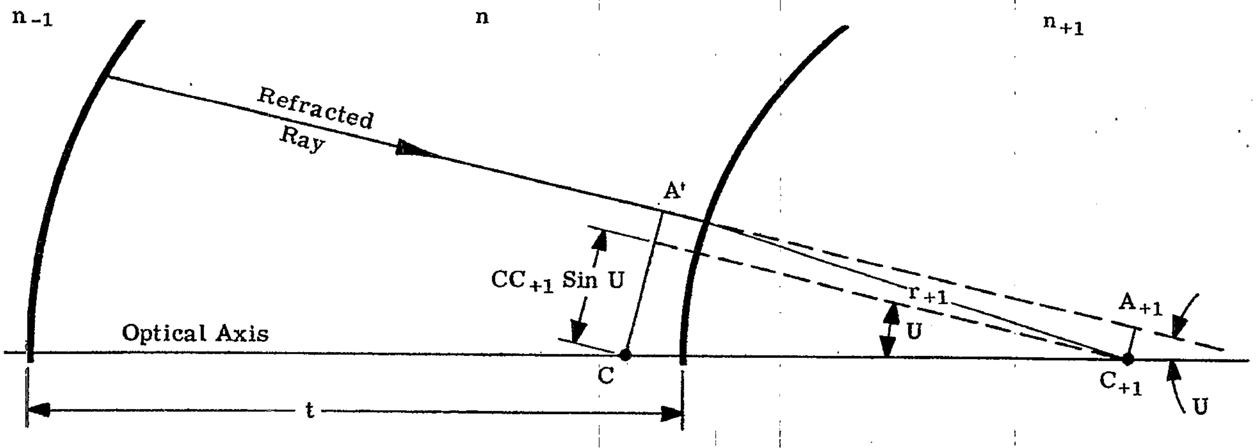


Figure 5.13 - Transfer procedure for the PR method.

5.6.5.4 A numerical example is shown in Table 5.5. The numbers above the double line are either given, such as r, t, and n, or are precalculated such as R, −R', and $(r - r_{+1} - t)\,n$. The P below the line, surface 1, is calculated from initial ray data, CA. All other values in the table, below the double line, are calculated using Equations (33) through (36). The problem of finding angles I and I' from their sines, in order to use Equation (36), is discussed below.

SURFACE		1		2		3
r		19.23		-64.25		13.51
t			0.8		0.05	
n	1		1.51017		1	
R		0.0520021		-0.0103063		
$-R'$		-0.0344346		0.0155642		
$n\ (r - r_{+1} - t)$			124.861		-77.81	
P		3.330000		10.708028		1.703612
$\sin I$		0.173167		-0.110360		
$-\sin I'$		-0.114667		0.166662		
U	0		-0.059124		-0.115983	

Table 5.5 - Numerical example of ray tracing by the PR method.

5.6.5.5 One should note that the above formula, (36), cannot be used to transfer from a plane surface wherein $r \rightarrow \infty$, or to a plane surface wherein $r_{+1} \rightarrow \infty$. To deal with a plane, the procedure is to calculate the distance from the pole of the plane surface to the ray; see Figure 5.14.

Let

$$OA\ n_{-1} = Q \text{ for the entering ray, and}$$

$$OA'\ n = Q' \text{ for the refracted ray.}$$

Then, because $U_{-1} = I$, and $U = I'$, we have,

$$Q' = Q\ \frac{\tan U_{-1}}{\tan U}\ .$$

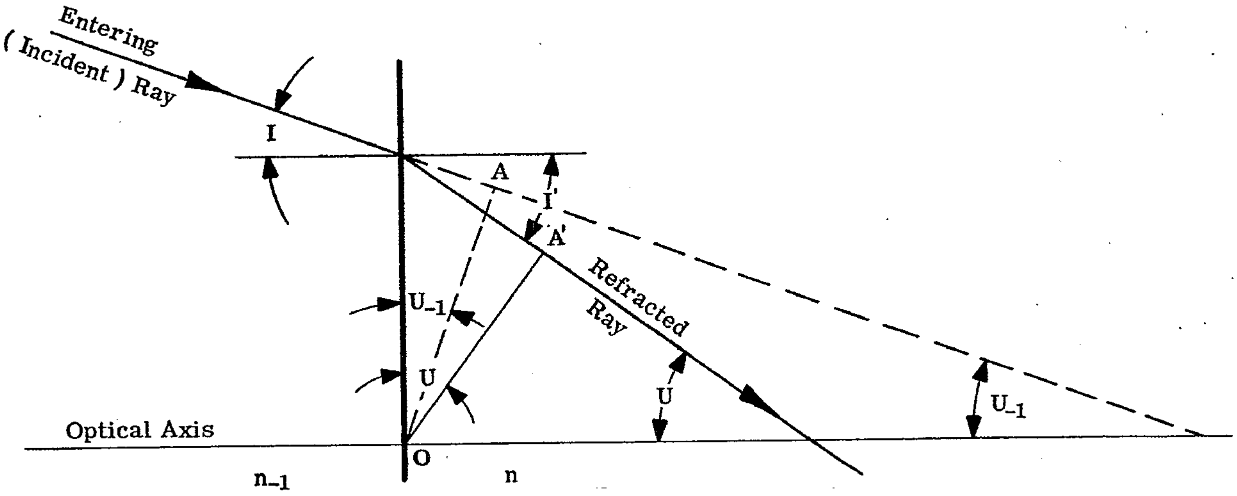


Figure 5.14 - Method of transfer for a plane surface.

To transfer from a spherical surface r, to a plane surface, $r_{+1} = \infty$, we use

$$Q_{+1} = P - (r - t)\, n \sin U .$$

To transfer from a plane surface, $r = \infty$, to a spherical surface, r_{+1}, the equation is

$$P_{+1} = Q' - (-r_{+1} - t)\, n \sin U .$$

5.6.5.6 To trace meridional rays through systems involving plane surfaces, Equations (33) through (36) are used until the plane surface is encountered. To transfer to a plane surface from a spherical one or vice versa, use one of the transfer equations in Paragraph 5.6.5.5. A transfer between two plane surfaces is calculated using

$$Q_{+1} = Q + t\; n \sin U .$$

Refraction at a plane surface is calculated using

$$\sin U = \frac{n_{-1}}{n} \sin U_{-1} ,$$

and

$$Q' = Q \frac{\tan U_{-1}}{\tan U} ,$$

where Q is either specified initially, calculated from initial data, or gotten by transfer from a previous surface. The calculations for plane surfaces are put into the same table (Table 5.5) as used for spherical surfaces. The values of $\sin U_{-1}$ and $\sin U$ are written opposite $\sin I$ and $\sin I'$ (which they equal respectively), and the values of Q and Q' are written opposite P. (The tangents need not be written down.)

5.6.5.7 One difficulty with the above formulation, Equation (36), is that if r_{+1} becomes large, but remains finite, P_{+1} becomes equal to the difference between a relatively small and a relatively large number. Hence unless a large number of significant figures are used for n, $\sin U$, and the coefficient of these terms, the value of P_{+1} will be independent of P. In doing hand computing one can readily notice this loss of precision. If this occurs, it is necessary to resort to other formulae, or reshape the lens so that the surface becomes plane. Another difficulty with Equation (36) arises if U becomes small, but remains finite. In this case the ray is almost parallel to the optical axis, and P_{+1} becomes equal to the difference of two nearly equal numbers. Hence unless both numbers are known to a large number of significant figures, the value of P_{+1} is quite inaccurate. In case the use of Equation (36) becomes difficult, the formulae given in Section 5.6.3 should be used.

5.6.5.8 In using the above equations it is necessary to convert sines to angles and to tangents, and to convert angles to sines. Tables are given in the Appendix. The tables convert from sine or tangent to the argument in radians and vice versa. They are designed for six place accuracy, and intervals are chosen for ease of interpolation. The first three digits of the function can always be found in the table and the last three digits are always multiplied by the interpolation constant and the product added to the tabular value. Interpolation therefore requires no mental arithmetic, and the process becomes completely automatic. By paying attention to such details a good human computer can trace rays through a lens at a speed of 40 to 60 seconds a surface. This method, in spite of its limitations, is an extremely useful method for hand computing meridional rays.

5.7 GRAPHICAL RAY TRACING PROCEDURE

5.7.1 Explanation of the method.

5.7.1.1 Rays may be traced graphically by means of a simple construction. The left side of Figure 5.15 shows a portion of two concentric circles whose radii are proportional to the indices n_{-1} and n. On the right side of the figure is shown the surface separating media of index n_{-1} and n. The angle of the refracted ray is determined from the diagram on the left. From this diagram $n_{-1} \sin I = n \sin I'$; thus, the construction solves Snell's law. Reference to Paragraph 5.4.5.3 will disclose that this is merely the graphical solution of the vector method.

5.7.1.2 The detailed procedure for tracing a ray is as follows. Draw a line through the center of the two circles parallel to the incident ray. Draw a line, parallel to the radius of curvature, through the intersection of the first line and the circle corresponding to the index of the object space. The line through the

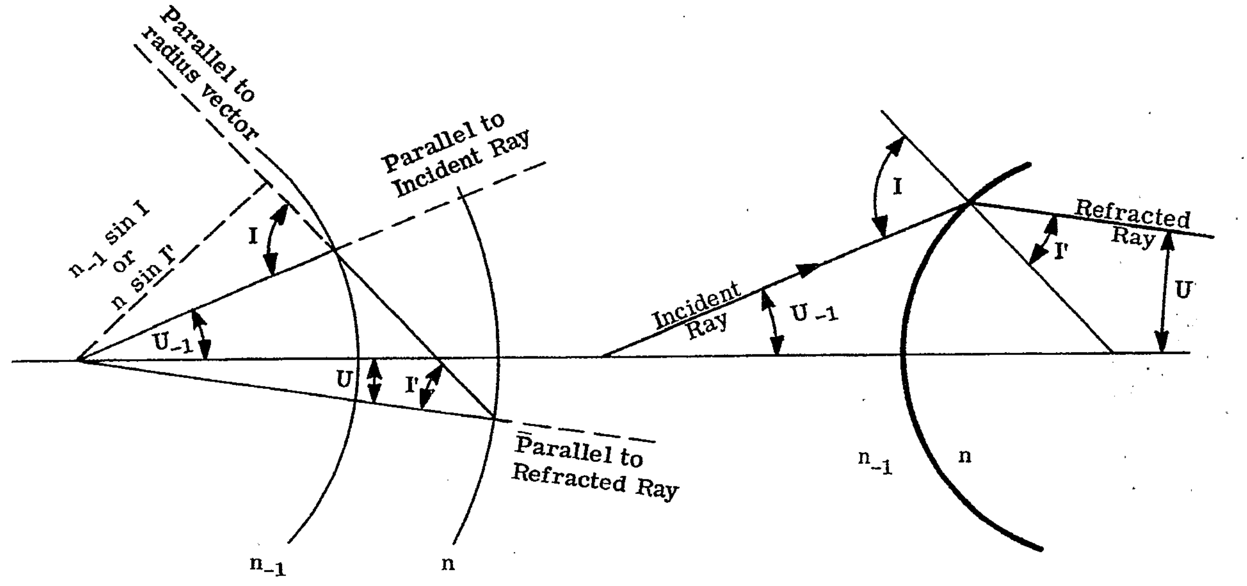


Figure 5. 15- The method of tracing rays graphically.

center of the two concentric circles and the intersection of the second line with the other circle is the refracted ray. The incident and refracted rays can be drawn on the right hand diagram, but this is not necessary. The two diagrams may be superposed by placing the center of the concentric circles on the incident ray a distance n_{-1} (arbitrary units) to the left of the incidence point. This procedure makes unnecessary the drawing of the circle for n_{-1} , or the drawing of two lines each for the incident ray and the radius vector. The remainder of the construction is as given above.

5.7.2 Example using an air-spaced doublet. Figure 5.16 shows the graphical ray trace for a ray which is initially parallel to the axis (ray a). It is seen that the first surface of the second lens is a diverging surface; the other three surfaces are converging, because the ray is bent toward the optical axis. By measurement of the radii of the concentric circles, we see that $n_1 = 1.5$ and $n_2 = 1.7$. This combination of a converging crown lens, followed by a diverging flint lens is typical of a type of achromatic telescope objective. These lenses will be studied in detail in Section 11 .

5.8 DIFFERENTIAL RAY TRACING PROCEDURE

5.8.1 Meaning of a differentially traced ray.

5.8.1.1 In the previous sections equations have been developed for tracing a general ray (skew or meridional) through a general surface having rotational symmetry. Once such a ray has been traced through the system, we have a baseline from which to find the path of neighboring rays. A differentially traced ray, sometimes referred to as a close ray, is a ray differing from the originally traced ray by small, first order quantities. This means that the change in direction cosines, dK_{-1} , dL_{-1} , dM_{-1} , and the change in the coordinates of the intersection point, dX , dY , dZ , are first order differentials.

5.8.1.2 The tracing of one ray gives us information about the one intersection point of that ray with the image surface. The tracing of several neighboring rays gives us their intersection points and hence information about the structure of the image formed by these rays. In addition to this useful information, differentially traced rays are generally easier to calculate than a single, general ray. Because of these advantages, the concepts and procedures of differential ray tracing are important.

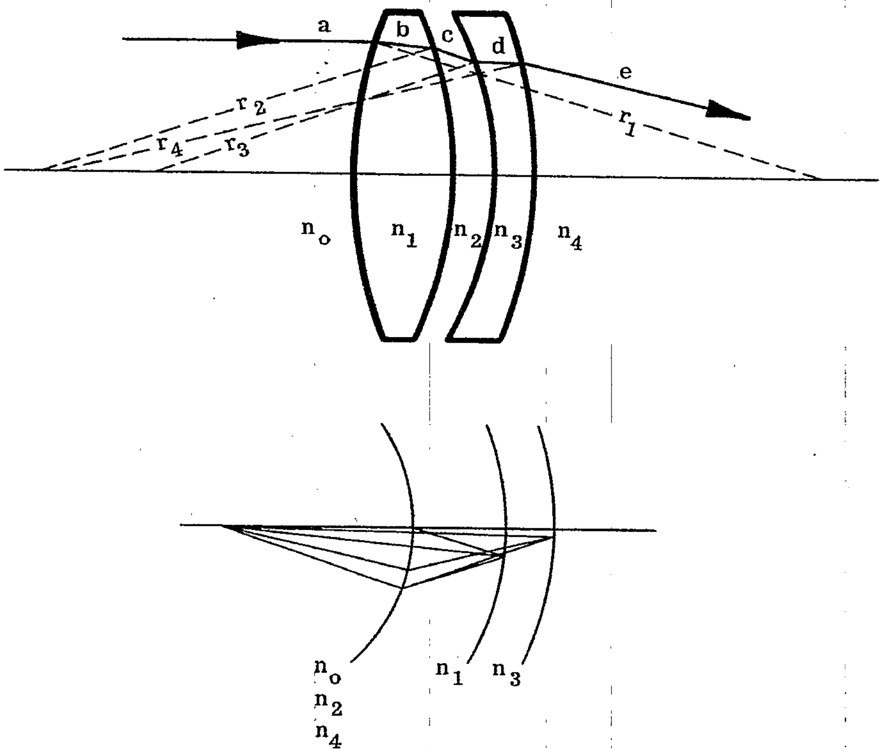


Figure 5.16 - Graphical ray trace of a doublet.

5.8.2 Differentially traced skew ray.

5.8.2.1 Once a skew ray has been traced through a lens system it is possible to trace the path of a ray differentially displaced from it. The skew ray trace provides the values of X, Y, Z on each surface, and K, L, M between surfaces. The values of X, Y, Z on adjacent surfaces are linked by the transfer equations

$$X = X_{-1} + \frac{D_{-1}}{n_{-1}} K_{-1} , \qquad (37)$$

$$Y = Y_{-1} + \frac{D_{-1}}{n_{-1}} L_{-1} , \qquad (38)$$

and

$$Z = Z_{-1} - t_{-1} + \frac{D_{-1}}{n_{-1}} M_{-1} , \qquad (39)$$

where D_{-1}, given by Equation (25a), is the geometrical distance along the skew ray between the two surfaces. These equations follow from Paragraph 5.4.3.2 applied to any two surfaces.

5.8.2.2 A neighboring ray, in the sense of Paragraph 5.8.1.1, will have slightly different coordinates on the jth surface. The differences, dX, dY, and dZ are found by differentiating Equations (37), (38), and (39). We have

$$dX = dX_{-1} + \frac{D_{-1}}{n_{-1}} dK_{-1} + K_{-1} \, d\left(\frac{D_{-1}}{n_{-1}}\right) , \qquad (40)$$

$$dY = dY_{-1} + \frac{D_{-1}}{n_{-1}} dL_{-1} + L_{-1} \, d\left(\frac{D_{-1}}{n_{-1}}\right) , \qquad (41)$$

and

$$dZ = dZ_{-1} + \frac{D_{-1}}{n_{-1}} dM_{-1} + M_{-1} \, d\left(\frac{D_{-1}}{n_{-1}}\right) . \qquad (42)$$

These equations may be referred to as differential transfer equations, in that they are used to calculate the change in coordinates. The changes in coordinates at the previous surface have been determined by the previous application of these equations; the changes in optical direction cosines are calculated by differential refraction equations discussed below. The last term, involving the change in total ray length, must also be calculated. The remaining equations will first be derived; then their order of use will be summarized.

5.8.2.3 The procedure used to derive an equation for $d\left(\frac{D_{-1}}{n_{-1}}\right)$ will be quite similar to that used to derive an equation for $\frac{A}{n_{-1}}$. (See Paragraphs 5.4.4.3 - 5.4.4.5). In that case we used four equations, Equations (4), (5), (6) and the equation for the sag, Paragraph 5.4.4.4. These four equations were solved simultaneously for $\frac{A}{n_{-1}}$. The equation for the sag, Z, is the equation for the surface, in that case the sphere. Because the intersection point must lie on the surface, this equation is called an equation of constraint. In the present case, the four equations to be used are Equations (40), (41), (42) and the differential equation of constraint.

5.8.2.4 Although the physical jth surface is a general surface of revolution, this surface is replaced by the plane, tangent at the intersection point. The reason this must be done is that we have restricted the change in coordinates to first order differentials; as one moves away from a point on a surface by distances of the order of first differentials, the motion is constrained to the plane tangent to the surface. The equation of the tangent plane is given in Paragraph 5.5.4.10. Differentiating this to obtain the differential equation of constraint we have

$$U\,dX + V\,dY + W\,dZ = 0 .$$

We now substitute into this equation the values of dX, dY, and dZ given by Equations (40), (41), and (42). Collecting terms, and using Equation (27), we have

$$d\left(\frac{D_{-1}}{n_{-1}}\right) = \frac{U\left(dX_{-1} + \frac{D_{-1}}{n_{-1}}\,dK_{-1}\right) + V\left(dY_{-1} + \frac{D_{-1}}{n_{-1}}\,dL_{-1}\right) + W\left(dZ_{-1} + \frac{D_{-1}}{n_{-1}}\,dM_{-1}\right)}{-G\,n_{-1}\,\cos I} . \tag{43}$$

5.8.2.5 Using Equation (43), and then Equations (40), (41), and (42), we will have completed the transfer of the differentially traced ray. The differential refraction equations, now to be derived, will be used to calculate dK, dL, and dM. Differentiating Equations (30) to (32) gives

$$dK = dK_{-1} + P\,dU + U\,dP , \tag{44}$$

$$dL = dL_{-1} + P\,dV + V\,dP , \tag{45}$$

and

$$dM = dM_{-1} + P\,dW + W\,dP . \tag{46}$$

5.8.2.6 In differentiating the equation for the tangent plane we kept U, V, and W constant and thereby obtained the differential equation of constraint. Physically this means that at any point on this tangent plane the ratio of the direction cosines of the normal, $U : V : W$, is the same as at any other point. (See Paragraph 5.5.5.3). Justifiably it may be asked why U, V, W were not held constant in differentiating Equations (30), (31), and (32). The answer is that though the tangent plane and surface differ by second order differentials, at the new intersection point, the normals to the two tangent planes, erected at the two intersection points, have direction cosines differing by first order differentials. Hence, since refraction involves the normal at the intersection point, dU, dV, and dW are not necessarily zero in Equations (44), (45), and (46).

5.8.2.7 Differentiating Equations (23), (24), and (18), we get

$$dU = -X\,dE - E\,dX , \tag{47}$$

$$dV = -Y\,dE - E\,dY , \tag{48}$$

and

$$dW = -\frac{c^2}{W}\left(X\,dX + Y\,dY\right) . \tag{49}$$

dE may be found by differentiating Equation (22), remembering that

$$dE = \frac{\partial E}{\partial X} dX + \frac{\partial E}{\partial Y} dY + \frac{\partial E}{\partial Z} dZ ,$$

thus

$$dE = -\left[\frac{c - E}{W} + \frac{2W^2}{c^2} \left(4e + 12fS^2 + 24gS^4 + 40hS^6 \right) \right] dW . \qquad (50)$$

5.8.2.8 The one remaining problem is the determination of dP to be used in Equations (44), (45), and (46). This is done by the same method used to derive Equation (43). The four equations used are Equations (44), (45), (46), and a differential equation of constraint. In each case, the fourth equation involves the differentials on the left-hand side of the first three equations. Because the sum of the squares of the direction cosines of a given line is unity, we have

$$KdK + LdL + MdM = 0$$

as the differential equation of constraint. Substituting Equations (44), (45), and (46) into this constraint, and remembering Equation (27), we have

$$dP = \frac{K(dK_{-1} + PdU) + L(dL_{-1} + PdV) + M(dM_{-1} + PdW)}{-G\, n \cos I'} . \qquad (51)$$

5.8.2.9 We can now summarize the calculations, in the order made, used in tracing a differentially traced skew ray through aspheric surfaces. In addition to the results for the skew ray, available from tables such as 5.3 and 5.4, the initial ray data of the neighboring ray must be given. That is dX_{-1}, dY_{-1}, dZ_{-1}, dK_{-1}, dL_{-1}, and dM_{-1} must be specified. The following equations are used in the order given here: (43), (40), (41), (42), (49), (50), (47), (48), (51), (44), (45), and (46). With an automatic computer it does not seem to be worthwhile to trace close skew rays since the regular skew rays can be traced so rapidly. However for hand computing the close skew ray trace is a very valuable tool.

5.8.3 Differentially traced meridional ray. At first sight it might appear that it would take as much time to trace a differentially traced ray as a skew ray. Actually the equations are simple and no square roots are involved. An interesting application of the above equations occurs in connection with meridional rays. Assume a meridional ray has been traced from an object point $(X_o = 0, Y_o, Z_o)$ through the lens system; let us now trace a ray from the same object point which will be differentially displaced by the amount dK_o. We also assume that $dL_o = 0$. Since

$$KdK + LdL + MdM = 0 ,$$

and the differential ray is to be traced around a meridional ray $K_o = 0$, $dM_o = 0$. (If originally we had assumed $dM_o = 0$, then it would follow that $dL_o = 0$). Equation (43) shows that $d(D_o/n_o) = 0$.

From Equations (40), (41) and (42),

$$dX_1 = \frac{D_o}{n_o} dK_o ,$$

$$dY_1 = 0 ,$$

and

$$dZ_1 = 0 .$$

Careful inspection of Equations (47) to (51) shows that,

$$dU = -EdX ,$$

$$dV = 0 ,$$

$$dW = 0 ,$$

$$dE = 0 ,$$

and

$$dP = 0 .$$

Substitution into Equation (44) gives the relation

$$dK_1 = dK_o - \left[\frac{G\,n_1 \cos I' - G\,n_o \cos I}{G^2} \right] E_1 \; dX_1 \,.$$

It then follows, since $d\left(\frac{D_1}{n_1}\right) = 0$, that

$$dX_2 = dX_1 + \frac{D_1}{n_1}\, dK_1 \,,$$

and

$$dK_2 = dK_1 - \left[\frac{G\,n_2 \cos I' - G\,n_1 \cos I}{G^2} \right] E_2 \; dX_2 \,.$$

The close meridional ray may be traced through the system by successive application of the equations

$$dX = dX_{-1} + \frac{D_{-1}}{n_{-1}}\, dK_{-1} \tag{52}$$

and

$$dK = dK_{-1} - \left[\frac{G\,n \cos I' - G\,n_{-1} \cos I}{G^2} \right] E\,dX \,. \tag{53}$$

5.8.4 The Coddington equations.

5.8.4.1 The above two equations,(52) and (53), apply to a general surface having rotational symmetry. In case the surface is spherical, Equation (53) is simplified since $E = c$ and $G = 1$. If the close ray has $dL_o = dM_o = 0$, as in the above example, and if the traced meridional ray and the close ray intersect to form an image, these two rays obey one of the Coddington equations, namely,

$$\frac{n_o}{D_o} + \frac{n_1}{D_1} = c\,(n_1 \cos I' - n_o \cos I) \,.$$

Because the close ray was shifted in a way that resulted in $dY_1 = dZ_1 = 0$, the shift of the intersection point occurred parallel to the X axis, in other words in the sagittal plane. The resulting focus is referred to as the sagittal focus, or the skew focus, because the close ray is actually a skew ray.

5.8.4.2 The above Coddington equation can be derived from Equations (52) and (53) applied to a spherical surface, $(E = c$ and $G = 1)$. Because we are dealing with a single object and single image point, $dX_o = dX_2 = 0$. Applying these two equations to Equation (52), we have

$$dX_1 = \frac{D_o}{n_o}\, dK_o = -\frac{D_1}{n_1}\, dK_1 \,.$$

Using Equation (53), for a spherical surface,

$$dK_1 = dK_o - (n_1 \cos I' - n_o \cos I)\, c\,dX_1 \,,$$

and, expressing dK_o in terms of dK_1, and dX_1 in terms of dK_1, we get

$$dK_1 = -\frac{D_1}{n_1}\frac{n_o}{D_o}\, dK_1 + (n_1 \cos I' - n_o \cos I)\, c\, \frac{D_1}{n_1}\, dK_1 \,.$$

Simplification gives the above Coddington equation.

5.8.4.3 Instead of shifting the ray in a plane perpendicular to the meridional plane, the ray could have been shifted in the meridional plane. In this case, $dK = dX = 0$. In a manner similar to that used in Section 5.8.3, ray trace equations for dY and dL can be derived, corresponding to Equations (52) and (53). (We do not need specific equations for dZ and dM, because these are proportional to dY and dL respectively). For a single image to be formed by two close rays from a single object point, $dY_o = dY_2 = 0$. The final result is the second Coddington equation involving the meridional or tangential focus,

$$\frac{n_o \cos^2 I}{D_o} + \frac{n_1 \cos^2 I'}{D_1} = c\,(n_1 \cos I' - n_o \cos I) \,.$$

5.9 PARAXIAL RAYS

5.9.1 The paraxial ray concept. The previous section on differentially traced meridional rays provides a good way to introduce the concept of paraxial ray tracing and the meaning of paraxial rays. A ray passing directly along the optical axis of the system is a perfectly good ray to use as a base from which to trace a close, neighboring ray. Such a ray, differentially traced with respect to the optical axis, is a paraxial ray. Physically, paraxial rays are the rays that get through the system as the aperture of each lens, centered concentrically with respect to the optical axis, becomes very small. Because paraxial rays are fairly easy to visualize, and because the ray tracing equations become quite simple for these rays, the usefulness of paraxial rays in the preliminary design of optical elements cannot be overemphasized.

5.9.2 Ray trace equations.

5.9.2.1 For a ray coinciding with the optical axis, $\cos I$ and $\cos I'$ will be exactly equal to 1 on every surface and $D_{-1} \rightarrow t_{-1}$, so Equations (52) and (53) become

$$dX = dX_{-1} + \frac{t_{-1}}{n_{-1}} dK_{-1} \qquad (54)$$

and

$$dK = dK_{-1} - (n - n_{-1}) c dX . \qquad (55)$$

Therefore a ray may be traced differentially close to the optical axis by applying the above equations. Since the original ray was the optical axis, there is no distinction between the X and Y axes, and these equations apply equally well for a close ray in the YZ plane. For such a ray, replace dX by dY and dK by dL , for each part of the system. It should be noted that these equations hold for aspheric as well as spherical surfaces. Mathematically this is so because for the optical axis, $X = Y = 0$; hence by Equations (18) and (26), $W = 1 = G$, and by Equation (22), $E = c$. Physically the aspheric and the sphere are tangent at the optical axis and have the same curvature; hence a ray close to the axis intersects a surface of curvature c .

5.9.2.2 Paraxial ray calculations will be used so extensively to build up an understanding of optical systems, that a special notation will be used to refer to paraxial data. It is customary to use lower case letters for paraxial rays. Equations (54) and (55) will be written for a ray in the YZ plane and become

$$y = y_{-1} + \frac{t_{-1}}{n_{-1}} (n_{-1} u_{-1}) , \qquad (56)$$

and

$$nu = n_{-1} u_{-1} + yc (n_{-1} - n) . \qquad (57)$$

The differentials have been replaced by small letters indicating paraxial ray data. One can see that dY has been replaced by y , indicating a small displacement perpendicular to the optical axis. dL , which replaces dK for a paraxial ray in the YZ plane, is the change in the optical direction cosine of the originally traced ray. Since the original ray is the axial ray, and the original $L = 0$, dL = new value of $L = n \cos \beta$, where β is the angle between the ray and the Y axis. Instead of $\cos \beta$, we can use $\sin U$, the angle between the ray and Z axis. Therefore $dL = n \sin U$. But U is a small angle, and we replace the $\sin U$ by U , the first order approximation. (See Section 5.11). Hence $dL = nU$, and using small letters, $dL = nu$. We see here why the term paraxial ray optics and first order optics are synonymous.

5.9.3 The use of finite angles and heights for paraxial rays.

5.9.3.1 Equations (56) and (57) were derived on the assumption that y and u are small, of the order of first order differentials. Physically, in order to form an image using paraxial rays, the actual rays must obey the condition that y and u are small. It is, however, both a remarkable and extremely useful fact that in ray tracing, we may use finite heights and angles, not necessarily small, for y and u . We will show this in the following paragraph.

5.9.3.2 Consider Figure 5.17 which indicates two rays from an axial object point O to the corresponding axial image point O'. Because u_{-1} and u in Equations (56) and (57) were assumed small, we can replace them by $\tan u_{-1}$ and $\tan u$ respectively. (The expansion of $\tan u$, in terms of u , shows that the first order approximation is $\tan u = u$, as in the case of $\sin u$. The third order approximation, how-

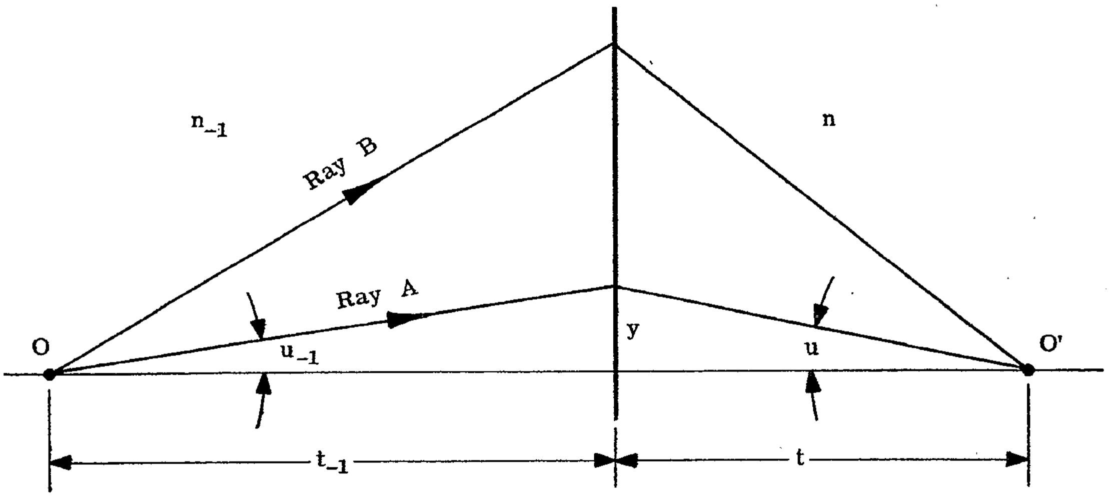


Figure 5.17. Paraxial rays through a single refracting surface.

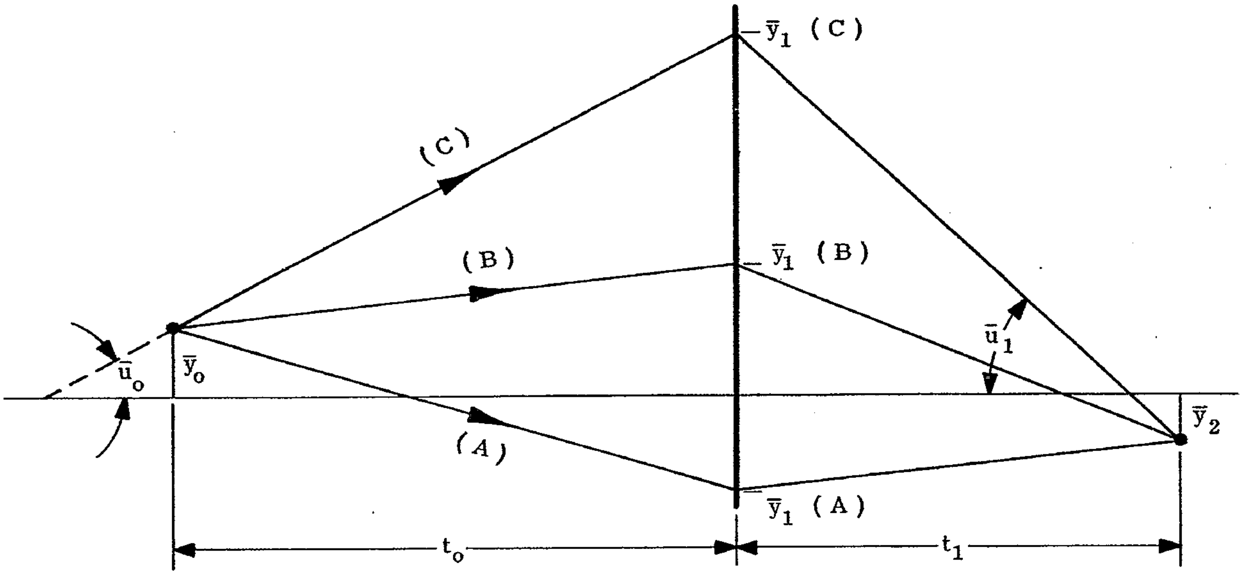


Figure 5.18 - Paraxial rays through a single refracting surface.

ever, differs from that of sin u). Remembering that u is negative, we substitute into Equation (57) and find

$$n \; -\!\left(\frac{y}{t}\right) = n_{-1}\,\frac{y}{t_{-1}} + \frac{y}{r}\,(n_{-1} - n)\,.$$

Upon rearrangement we get

$$\frac{n_{-1}}{t_{-1}} + \frac{n}{t} = \frac{n - n_{-1}}{r}\,,$$

which is the familiar form of the paraxial equation for a single surface. The important thing to note is that u_{-1} , u , and y no longer appear in this equation. This fact is interpreted as meaning that mathematically we may consider the image O' formed by any ray leaving the object O . Thus both rays (A) and (B) intersect the axis at the same image point.

5.9.3.3 Figure 5.17 and the above paragraph apply to axial object and image points. The same conclusions concerning finite heights and angles hold for rays through off-axis object and image points. Hence, in Figure 5.18, all rays (A), (B), and (C) intersect at one image point. Neither the angles $\bar{u}_o$ or $\bar{u}_1$, nor the heights $\bar{y}_0$, $\bar{y}_1$, or $\bar{y}_2$, need be small.*

5.10 GRAPHICAL RAY TRACE FOR PARAXIAL RAYS

5.10.1 Specialization of the general graphical method.

5.10.1.1 Paraxial rays may be traced graphically through a lens system by a construction very similar to the construction shown in Section 5.7. This is done by replacing the refractive index circles by tangent planes, and the curves of the lens surfaces by tangent planes through the vertices of the surfaces. The justification for these replacements will be given in Paragraph 5.10.1.3. For paraxial rays, the construction will appear as shown in Figure 5.19.

5.10.1.2 In the above paragraph we have indicated that Figure 5.19 is correct for paraxial rays.

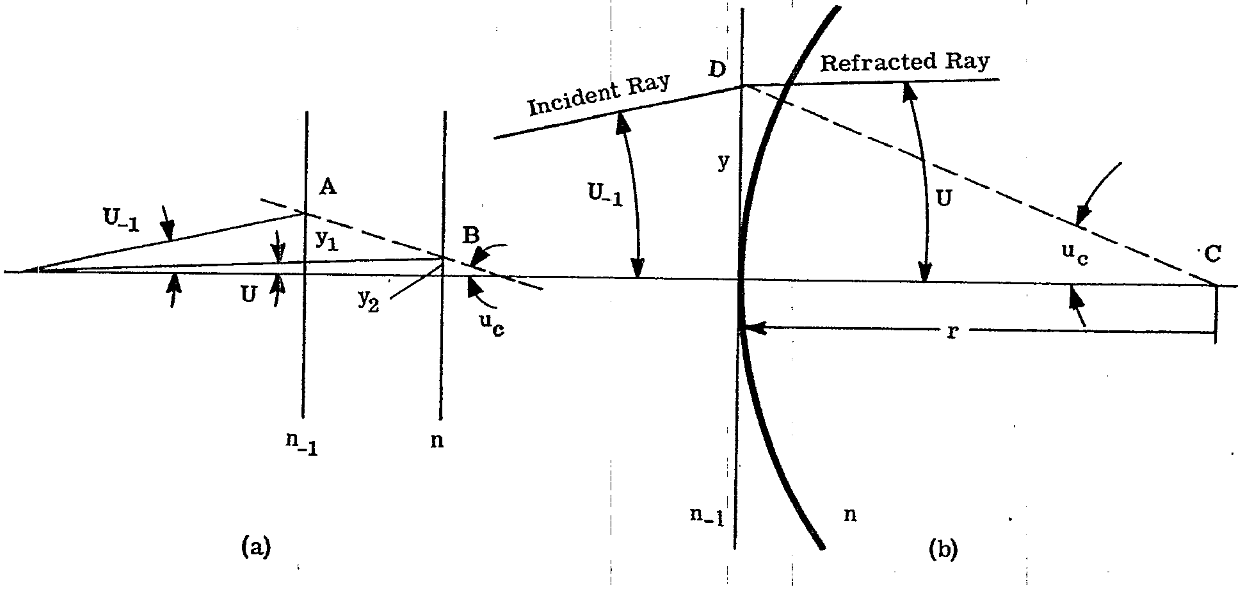


Figure 5.19 - The method for tracing paraxial rays graphically.

---

* The angles and heights corresponding to rays through off-axis object and image points are written with a line or bar over the symbol, as $\bar{u}$ and $\bar{y}$.

Assuming this let us use the drawing to reexamine and extend the ideas discussed in Section 5.9.3. From Figure 5.19 it can be seen that

$$y_1 = n_{-1} \tan U_{-1}$$

and

$$y_2 = n \tan U .$$

Since the line connecting A B (a) is parallel to line DC (b) it is clear that, from similar triangles,

$$\frac{y_1 - y_2}{y} = \frac{n - n_{-1}}{r} .$$

By inserting the expressions for y_1 and y_2 into the last equation, we find on rearranging

$$n \tan U = n_{-1} \tan U_{-1} + yc\,(n_{-1} - n) .$$

5.10.1.3 This last equation, derived from Figure 5.19, is correct for small angles. This is easily seen, because when the $\tan U_{-1}$ and $\tan U$ are replaced by the angles u_{-1} and u respectively, Equation (57) results. However let us assume for the moment that both Figure 5.19 and the last equation are correct for any angles U and U_{-1}. In particular, these may be finite angles and do not have to be small. Now if we compare this equation to Equation (57), which is true for small angles u and u_{-1}, and therefore for paraxial rays, we see that $\tan U$ and $\tan U_{-1}$ correspond to u and u_{-1}, respectively. This indicates that the equation derived from Figure 5.19 can be used in connection with paraxial rays, provided the angles u and u_{-1} are replaced by $\tan U$ and $\tan U_{-1}$ respectively. Since U can have any value, $\tan U$ and therefore u can have any value. Equation (57) and Figure 5.19 can therefore be used for paraxial rays incident at any finite height and making any finite angle with the optical axis. Equation (56) and Figure 5.19 can also be used to accurately transfer the value of y from one surface to another for paraxial rays. This equation and figure can also be used with non-paraxial meridional rays to transfer between plane surfaces; in this case u_{-1} is replaced by $\tan U_{-1}$.

5.10.2 Two approaches to the treatment of paraxial rays.

5.10.2.1 We have shown that paraxial rays can be considered from either of two points of view:

(1) We use small angles and finite curvatures for surfaces. This led to Equation (57).

(2) We use finite angles and zero curvatures for surfaces. This led to Figure 5.19. It must be emphasized that we do not have to combine these and use small angles with plane surfaces.

5.10.2.2 It is convenient then to think of paraxial rays as passing through the optical system at finite heights, striking the surfaces on the tangent planes instead of the actual curved surfaces. Since the two Equations (56) and (57) are linear equations, and since the location of images are found for values of $y_k = 0$, it makes no difference what value of u is used. It is instructive to trace paraxial rays through a lens at heights equal to the actual ray heights, and note the difference in path for a paraxial ray and an actual ray. This is demonstrated for a single surface refraction in Figure 5.20. The ray traced through the curved surface crosses the axis at M, closer to the surface than the point P. The paraxial ray crosses at P, further away from the surface. This defect of focus is called spherical aberration.

5.11 THE DIFFERENT "ORDERS" OF OPTICS

5.11.1 Expansion of the sine function.

5.11.1.1 The fundamental equations which have been discussed and used in tracing rays are: (1) the transfer equations, and (2) the refraction equations. Both have been put into a form explicitly using the cosine function of various angles, such as the angles of incidence and refraction, and the angles which the ray makes with the coordinate axes. Both equations could have been written in terms of the sine function; so as to explain the meaning of the phrase orders of optics we will deal with the sine function.

5.11.1.2 The optical axis is a special ray for which both angles of incidence and refraction are zero. In addition the angles which this ray makes with the X, Y, and Z axes are 90°, 90°, and 0° respectively. For a meridional ray near the axis, the angles of incidence and refraction, and the angle with the Z axis, are small. The ray trace equations, therefore, involve the sines of small angles. As the meridional ray

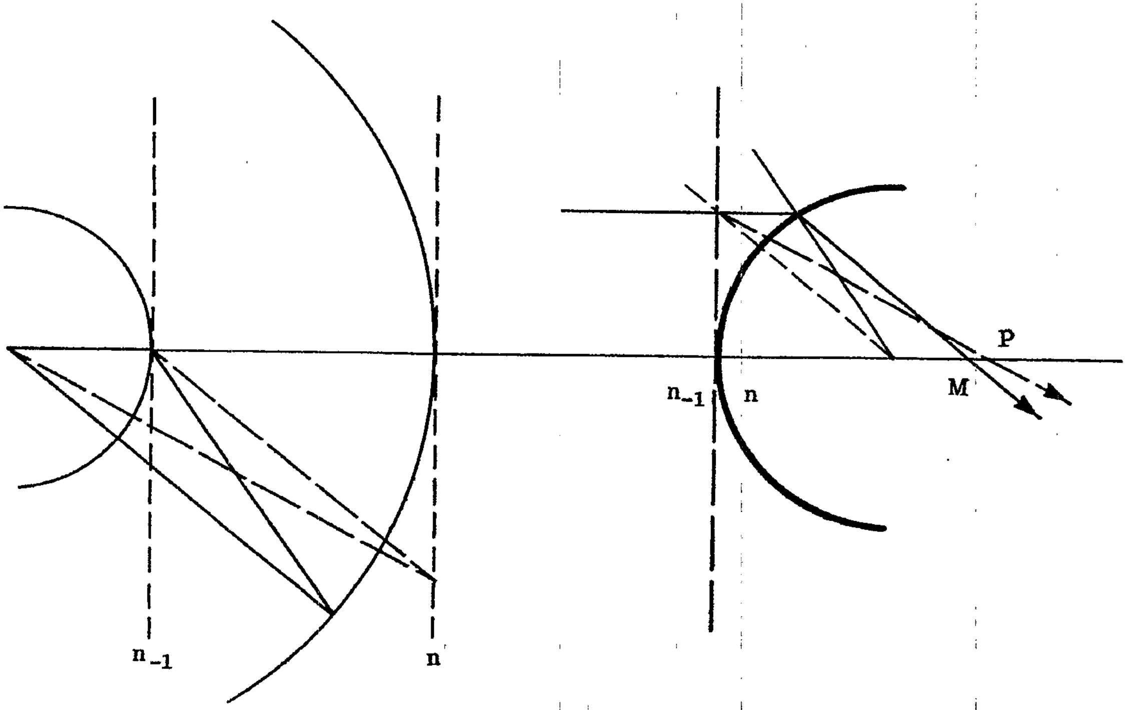


Figure 5.20 - Comparison between a paraxial and an actual ray showing spherical aberration.

makes larger and larger angles with the Z axis, we have to be concerned with the sines of larger and larger angles.

5.11.1.3 One reason the ray trace equations are complicated is that they involve the trigonometric functions of angles, instead of just the angles. (We have seen in Section 5.9 how the equations are greatly simplified when they can be expressed in terms of angles, instead of trigonometric functions). To relate the $\sin \alpha$ to the angle α, we expand the sine function in a series, thus

$$\sin \alpha = 0 + \alpha - 0 - \frac{\alpha^3}{3!} + 0 + \frac{\alpha^5}{5!} - 0 - \frac{\alpha^7}{7!} + \ldots$$

5.11.1.4 The terms given explicitly as zero have been written down to clarify the situation. Whenever a function is expanded in a series the "first" term is called the zeroth approximation or the zeroth order, and successive terms are called the first, second, third, etc., orders. In the case of the expansion of the sine function, the zeroth, second, fourth, and all even order terms are identically zero; only the odd orders remain.

5.11.2 First order optics. If in ray trace equations the sine is replaced by the angle, we are using the zeroth and the first order terms in the above expansion. The resulting equations and design procedures are called first order optics, and the rays concerned are paraxial rays. One of the fascinating parts of geometrical optics is the extensive understanding of lens systems one can obtain by tracing two paraxial rays. With two paraxial rays one can predict the location and size of any image formed with paraxial rays, and by making further calculations based on these paraxial ray data it is possible to predict the approximate magnitude of image errors. The following sections, 6 and 7, will be devoted to the development and use of the equations of first order optics. This development will be based on the two simple equations, (56) and (57).

5.11.3 Third order optics.

5.11.3.1 If the first and third order terms in the expansion of the sine are retained, the resulting equations are part of third order optics. But this term has an added meaning, pertaining to aberrations, and it is usually in this latter sense that the term is used.

5.11.3.2 The intersection of a ray with the image surface locates the image. If the intersection has been computed using the skew ray trace equations, the intersection is the true image. If paraxial ray trace equations have been used, the resulting paraxial image will generally be displaced from the true image. The difference between the true image and the first order approximation (paraxial approximation) is known by the general term aberration. (We are considering here monochromatic light only. Aberrations due to non-monochromatic light are considered in Paragraph 5.11.3.4.) In the same way that the sine was expanded in a series, the aberrations can be expanded. The first term in the expansion is known as the third order aberration. The reason for this is that it represents the first approximation to the total aberration, and hence can be considered as the difference between the paraxial image and the image using the third order approximation for the sine. Third order optics then has come to mean the equations and procedures dealing with the first approximation to the aberrations. It is fortunate, as will be evident in a later section (Section 8) that these third order aberrations can be calculated from first order (paraxial) ray trace data.

5.11.3.3 The next term in the expansion of the aberration, after the third order aberration, is called the fifth order aberration. Fifth order optics deals with the aberrations through the fifth order aberration term.* Hence fifth order optics deals with fifth order aberration, or the second approximation to the aberration.

5.11.3.4 Aberrations due to non-monochromatic light can also be expanded in a series. The first term gives the aberration appearing in paraxial images, hence is referred to as first order aberration. This is treated in Section 6, dealing with first order optics.

---

* In some countries other than the United States, for example England, the first, second, third, etc., terms in the aberration expansion are referred to as primary, secondary, tertiary, etc. aberration.

6 FIRST ORDER OPTICS

6.1 GENERAL

6.1.1 First order optics and paraxial rays. In Section 5.11.2 it was pointed out that when the sine of the angle is replaced by the angle, the resulting equations belong to the field of first order optics. In general, if any trigonometric function is replaced by its first approximation, we get first order equations, in the field of optics. In Sections 5.9.1 and 5.9.2 we defined a paraxial ray as one differentially displaced from the optical axis. Because of this definition we must use the first approximation to the trigonometric functions in the equations for a differentially traced ray. The resulting paraxial ray equations are hence identical to the first order equations.

6.1.2 Preliminary layout and graphical ray trace. The method of tracing paraxial rays graphically was explained in Section 5.10. Graphical ray tracing is extremely useful in the preliminary design stage, particularly for complicated systems, which cannot be visualized easily. The designer can thereby get a "feel" for the system, which a mere array of numbers often hides. Graphical ray tracing, however, is limited to an accuracy of about one percent. For additional accuracy, which is absolutely necessary in the calculation of aberrations, we must resort to numerical paraxial ray tracing. The methods and results of this type of ray tracing in the realm of first order optics will be discussed in Section 6.

6.2 NUMERICAL TRACING OF PARAXIAL RAYS

6.2.1 Importance of paraxial ray tracing. The accurate numerical tracing of paraxial rays is used extensively in the design of optical systems for three main reasons:

(1) Tracing paraxial rays through the system is a simple mathematical procedure.

(2) Images formed by paraxial rays provide very convenient reference planes.

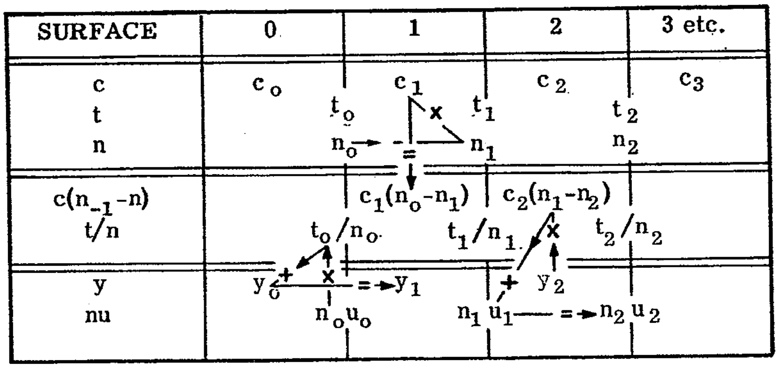
(3) Data obtained in paraxial ray calculations can be used to calculate the first approximation to image aberration.

For these reasons, a systematic method of numerical ray tracing of paraxial rays is a necessary tool for the designer, even today when large automatic computers are readily available. In this section such a method will be described; and it will be used extensively in the following sections to illustrate the vast amount of information made available by paraxial ray tracing.

6.2.2 Ray trace format.

6.2.2.1 The first step in tracing a paraxial ray is to lay out the system data in a form as shown in the top of Table 5.1. Then the two constants, $c\,(n_{-1} - n)$ and t/n, are computed for each surface and space respectively. (See Table 6.1). With these constants filled in, the paraxial ray may be traced by applying Equations (56) and (57) of Section 5.

6.2.2.2 As one uses this representation, its value becomes evident. These equations, and the way the data



SURFACE	0	1	2	3 etc.
c	c_0	c_1	c_2	c_3
t	t_0	t_1	t_2	
n	n_0	n_1	n_2	
$c(n_{-1}-n)$		$c_1(n_0-n_1)$	$c_2(n_1-n_2)$	
t/n	t_0/n_0	t_1/n_1	t_2/n_2	
y	y_0	y_1	y_2	
nu	n_0u_0	n_1u_1	n_2u_2	

Table 6.1 - Recommended format for tracing paraxial rays through an optical system.

are laid out, make almost a perfect match with the requirements of a desk calculator. A few of these features are:

(1) In calculating $c(n_{-1} - n)$ one obtains the data from a triangle of numbers,

(2) In tracing the ray, both equations are computed in the same way. First a number is multiplied by a number directly above it, then the product added to the number below the double line on the left, and the result written in the space on the right. This is indicated by the lines shown in the figure that appear as and

(3) Many times, problems are worked backwards. For example, suppose $n_{-1}\ u_{-1}$, nu, and y are given, and the problem is to find c, The question is: how to remember what to do first, i.e., divide y by $(nu - n_{-1}\ u_{-1})$, or vice versa? It turns out that the correct method is always the easiest one to do on the calculating machine. Dividing $(nu - n_{-1}\ u_{-1})$ by y can be done without writing down $(nu - n_{-1}\ u_{-1})$. However, to calculate $y/(nu - n_{-1}\ u_{-1})$, the difference must be written down; therefore, we know that to calculate c, the result must be $(nu - n_{-1}\ u_{-1})/y$ divided by $(n_{-1} - n)$. As another example, suppose a value of y_1, y_2, and $n_1\ u_1$ are given, what t_1/n_1 is needed? The formula can be remembered in the following way: first compute $y_2 - y_1$, and then divide by $n_1\ u_1$. Therefore the formula is $t_1/n_1 = (y_2 - y_1)/n_1\ u_1$.

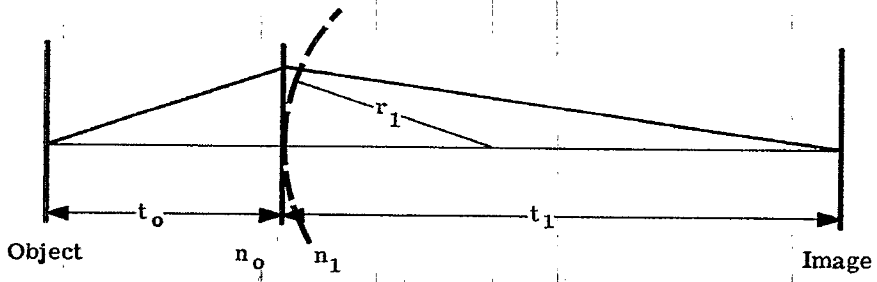


Figure 6.1 - Relation between image and object points.

6.2.3 Algebraic example. Table 6.1 may also be used to derive algebraic expressions useful in optics. One can very readily work out the equation for image and object distances for a single refracting surface. If a surface of radius r_1 separates two media of index n_o and n_1, an object point will be imaged at a distance t_1 from the surface (see Figure 6.1). What is the relation between t_o and t_1? This is a three surface problem, the object surface, 0, the refracting surface, 1, and the image surface, 2. A paraxial ray at $y_o = 0$ will be imaged at $y_2 = 0$. It is therefore possible to fill out the calculations of Table 6.1 to the following extent. (See Table 6.2).

SURFACE	OBJECT	1	IMAGE
c	0	c_1	0
t	t_o	t_1	
n	n_o	n_1	
$c(n_{-1} - n)$	0	$c_1(n_o - n_1)$	0
t/n	t_o/n_o	t_1/n_1	
y	0		0
nu			

Table 6.2 - Single refracting surface, axial object and image points.

Now, as pointed out in Sections 5.9.3 and 5.10.1, the angle used to trace a paraxial ray does not affect the image position. Since the choice of y on the lens is arbitrary, let it be y_1 . The calculations to this point are shown in Table 6.3.

$c(n_{-1}-n)$	0		$c_1(n_o-n_1)$		0
t/n		t_o/n_o		t_1/n_1	
y	0		y_1		0
nu		$n_o u_o$		$n_1 u_1$	

Table 6.3 – Continuation of Table 6.2.

With y_o (= 0) and y_1 (arbitrary) known, $n_o u_o = (y_1 - 0)/(t_o/n_o)$.

With $n_o u_o$ and y_1 known, $n_1 u_1 = n_o u_o + y_1 c_1 (n_o - n_1)$.

With $n_1 u_1$, y_1 and y_2 (= 0) known, $t_1/n_1 = (0 - y_1)/n_1 u_1$.

Therefore $t_1/n_1 = -y_1/n_1 u_1$, or

$$\frac{n_1}{t_1} = -\frac{n_1 u_1}{y_1} = \frac{-n_o u_o}{y_1} - c_1 (n_o - n_1) = -\frac{n_o}{t_o} + c_1 (n_1 - n_o).$$

This equation becomes the familiar refraction equation, derived in Paragraph 5.9.3.2,

$$\frac{n_1}{t_1} + \frac{n_o}{t_o} = c_1 (n_1 - n_o) . \tag{1}$$

Notice how the y_1 has dropped out of the equation indicating that any value y_1 could have been used. The calculations will be finally filled out in Table 6.4 as follows:

y	0		y_1		0
nu		$y_1 n_o/t_o$		$y_1 n_o/t_o$ $+ y_1 c_1(n_o-n_1)$	

Table 6.4–Conclusion of Table 6.3.

6.2.4 Numerical example. Equation (1) was given to show how the ray trace table can be used to derive a classical formula. Actually one will find very little occasion to use Equation (1) to calculate a numerical result, because problems can be solved much more readily using the format of Table 6.1. For example, suppose one is given the problem $c_1 = 0.10$, $n_o = 1$, $n_1 = 1.5$, $t_o = 10$. Rather than remember any special formula, go directly to the format as shown in Table 6.5 .

SURFACE	OBJECT		1		IMAGE
c	0		0.10		0
t		10		t_1	
n		1		1.5	
$c(n_{-1}-n)$	0		-0.05		0
t/n		10		$t_1/1.5$	
y	0		1		0
nu		0.10		0.05	

$$(0.05)\frac{t_1}{1.5} + 1 = 0$$

$$\frac{t_1}{1.5} = \frac{-1}{.05} = -20$$

$$t_1 = (1.5)\,(-20) = -30$$

Table 6.5–Numerical example of a single refracting surface.

6.2.5 Ray trace for three element lens. Table 6.6 shows the data and ray trace results for a three element lens. All the material above the lowest double line has been discussed earlier in this chapter. The last two lines, involving $\bar{y}$ and $n\bar{u}$, and the calculations of m (lateral magnification), f' (focal length) , and Φ (optical invariant) will be discussed in the following sections.

SURFACE	OBJECT 0	1	2	3	4	5	6	IMAGE 7
c	0	0.25285	-0.01474	-0.19942	0.25973	0.05065	-0.24588	0
t	25.00000	0.60000	1.06541	0.15000	1.13691	0.60000	14.05015	
n	1.00000	1.62000	1.00000	1.62100	1.00000	1.62000	1.00000	
$c(n_{-1} - n)$		-0.15677	-0.00914	0.12384	0.16129	-0.03140	-0.15245	
t/n	25.00000	0.37037	1.06541	0.09254	1.13691	0.37037	14.05015	
y	0	1.25000	1.19594	1.02879	1.02606	1.18070	1.21734	0
nu	0.05000	-0.14596	-0.15689	-0.02948	0.13601	0.09894	-0.08664	
$\bar{y}$	-10.00000	-0.75000	-0.56942	-0.04440	0.00069	0.55481	0.72887	5.77084
$n\bar{u}$	0.37000	0.48758	0.49278	0.48728	0.48739	0.46997	0.35886	

$$m = \frac{n_o u_o}{n_6 u_6} = -0.57708 = \frac{\bar{y}_7}{\bar{y}_o} \qquad \Phi = -0.50000$$

$$f' = -\frac{\Phi}{n_o (u_o \bar{u}_6 - \bar{u}_o u_6)} = \frac{0.5}{1\ (0.05 \times 0.35886 + 0.37 \times 0.08664)} = 10.000$$

Table 6.6 Sample calculation of paraxial rays through a three element lens, using Equations 5-(56) and 5-(57).

6.2.6 Ray trace procedure for calculation of aberrations. Another way to trace paraxial rays is to use the following equations:

$$y = y_{-1} + t_{-1} u_{-1} \tag{2}$$

$$u = u_{-1} + i \left[(n_{-1}/n) - 1 \right] \tag{3}$$

This ray trace involves the new quantity, i , which is the limiting value of the angle of incidence, I , as the ray approaches the axis in the paraxial region. Equation (2) is merely Equation 5-(56) simplified. Equation (3) comes from Equation 5-(35), written for small angles, with the substitution $i' = i\ n_{-1} / n$; the latter is the law of refraction for small angles. Now from Figure 5.11, $I' - U$ = the acute angle between r and the optical axis. But for small angles this is $y / r = y\,c$. Hence, using Equation 5-(35) we have,

$$i = y\,c + u_{-1} \;. \tag{4}$$

It will be shown later (Section 8) how the third order aberrations may be calculated from paraxial ray data. For these calculations it is easier to use Equations (2), (3) and (4) ,than Equations 5-(56) and 5-(57).

6.2.7 Numerical example.

6.2.7.1 To illustrate these equations, Table 6.7 includes paraxial rays traced through the same lens as used in Table 6.6. In this example, a different set of rays are traced through the lens. Below the lowest double line there are entries used in the calculation of chromatic aberration. These calculations will be explained in Section 6.10.

SURFACE	OBJECT 0	1	2	3	4	5	6	IMAGE 7
c	0	0.252850	-0.014740	-0.199420	0.259730	0.050650	-0.245880	
t	∞	0.600000	1.065410	0.150000	1.136910	0.600000	8.279369	
n	1.000000	1.620000	1.000000	1.621000	1.000000	1.620000	1.000000	
$(n_{-1} / n) - 1$		-0.382716	0.620000	-0.383097	0.621000	-0.382716	0.620000	
y	0	1.500000	1.412907	1.148619	1.138827	1.227353	1.241918	0
u	0	-0.145155	-0.248063	-0.065279	0.077866	0.024274	-0.150001	
i		0.379275	-0.165981	-0.477120	0.230508	0.140031	-0.281089	
dn/n	0	0.006370	0	0.010586	0	0.006370		TAch = -0.0029
$\Delta(dn/n)$		0.006370	-0.006370	0.010586	-0.010586	0.006370	-0.006370	
$a = -y n_{-1}\ i \Delta \frac{dn}{n}$		-0.00362	-0.00242	0.00580	0.00450	-0.00109	-0.00360	$\Sigma a = -0.00044$

Table 6.7 - A paraxial ray is traced through the same lens as used in Table 6.6. In this case Equations (2), (3), and (4) are used.

6.2.7.2 For use with a large computing machine there is no preference for either of these methods. For hand computing, unless aberrations are calculated, the method outlined in Table 6.1 is simpler. Therefore, all the paraxial ray theory given in Section 6.3-6.9 will be based on Equations 5-(56) and 5-(57).

6.3 THE OPTICAL INVARIANT

6.3.1 Axial and oblique rays. In Section 6.2 it was shown how images may be located along the axis of the optical system. The procedure is to trace a paraxial ray from where the object surface crosses the optical axis ($y_o = 0$). Such a ray is called an axial paraxial ray. An image surface is formed wherever this paraxial ray crosses the optical axis. By tracing a second ray from the object at a value of $y_o \neq 0$ it is possible also to determine the size of the image. Such a ray is called an oblique paraxial ray. The data for this second ray will be identified by writing $\bar{y}$ and $\bar{u}$. Table 6.6 shows a second ray traced through the lens. The second ray is commonly referred to as the oblique paraxial ray because it passes from an off-axis object point obliquely through the optical system to the image. If this ray passes through the center of the aperture stop it is called a chief ray. In tracing the oblique paraxial and the axial paraxial ray through the system, the following equations have been applied for each surface:

$$nu = n_{-1} u_{-1} + yc(n_{-1} - n) \quad \text{for the axial paraxial ray refraction.} \qquad 5\text{-}(57a)$$

$$n\bar{u} = n_{-1} \bar{u}_{-1} + \bar{y}c(n_{-1} - n) \quad \text{for the oblique paraxial ray refraction.} \qquad 5\text{-}(57b)$$

$$y = y_{-1} + \frac{t_{-1}}{n_{-1}}(n_{-1} u_{-1}) \quad \text{for the axial paraxial ray transfer.} \qquad 5\text{-}(56a)$$

$$\bar{y} = \bar{y}_{-1} + \frac{t_{-1}}{n_{-1}}(n_{-1} \bar{u}_{-1}) \quad \text{for the oblique paraxial ray transfer.} \qquad 5\text{-}(56b)$$

6.3.2 The optical invariant and its importance. We will use the last four equations, involving axial and oblique paraxial rays, to derive an expression called the optical invariant. This quantity, as its name implies, is a constant; as such it may be calculated in several ways and its value for a given system can be used in the calculation of various quantities. This invariant has a meaning for an optical system similar to momentum or energy for an isolated mechanical system.

6.3.3 The invariant for refraction. By transposition and division, using Equations 5-(57a) and 5-(57b), it is possible to equate the common term $c(n_{-1} - n)$ giving

$$\frac{nu - n_{-1} u_{-1}}{y} = \frac{n\bar{u} - n_{-1} \bar{u}_{-1}}{\bar{y}} .$$

By rearranging, this may be written

$$\bar{y}(n_{-1} u_{-1}) - y(n_{-1} \bar{u}_{-1}) = \bar{y}(nu) - y(n\bar{u}) . \qquad (5)$$

The index and angle data on the left side of this equation refer to the space to the left of the surface, and the corresponding data on the right side refer to the space to the right of the surface. This equation shows that

$$\bar{y}(nu) - y(n\bar{u}) = \Phi \qquad (6)$$

is an invariant for the refraction at any surface in the optical system. Φ is called the optical invariant.

6.3.4 The invariant for transfer. In a similar way Equations 5-(56a) and 5-(56b) may be combined to give the relation

$$\bar{y}_{-1}(n_{-1} u_{-1}) - y_{-1}(n_{-1} \bar{u}_{-1}) = \bar{y}(n_{-1} u_{-1}) - y(n_{-1} \bar{u}_{-1}) .$$

It is noted that the right hand side of this equation is equal to the left hand side of Equation (5), and hence is Φ, the optical invariant. Moreover both y values on the left apply to the surface to the left of the space, and both y values on the right refer to the surface to the right of the space. Therefore this equation shows that the optical invariant is also an invariant as the ray is transferred from one surface to the next.

6.3.5 The invariant for the entire system. We have shown above that there is a combination of y , n , u , $\bar{y}$, and $\bar{u}$, which has the same value on either side of a surface, that is, it is invariant across a surface between two spaces. We have also shown that this same combination of parameters is the same on either side of a space, that is, it is invariant across a space between two surfaces. Hence the optical invariant is an invariant for an entire optical system. It is therefore possible to write down the optical invariant between any two surfaces (or any two spaces). For example, between the object surface and the image surface we can write

$$\Phi = \bar{y}_o (n_o u_o) - y_o (n_o \bar{u}_o) = \bar{y}_k (n_{k-1} u_{k-1}) - y_k (n_{k-1} \bar{u}_{k-1}) .$$

The invariant may also be written in determinant form as

$$\Phi = \begin{vmatrix} \bar{y} & n\bar{u} \\ y & nu \end{vmatrix} .$$

6.3.6 Lateral magnification. If $y_o = 0$ on the object surface (the 0th surface), and $y_k = 0$ on the image surface (the kth surface), then the next to the last equation becomes

$$\Phi = \bar{y}_o (n_o u_o) = \bar{y}_k (n_{k-1} u_{k-1}) .$$

This is illustrated in Figure 6.2.

Using the optical invariant then, it is possible to calculate the height of the image $\bar{y}_k$ from the object height, $\bar{y}_o$. The lateral magnification, m, is defined as

$$m = \frac{\bar{y}_k}{\bar{y}_o} = \frac{(n_o u_o)}{(n_{k-1} u_{k-1})} . \qquad (7)$$

This equation shows that the lateral magnification can be calculated by tracing a single paraxial ray from the base of an object to the base of the image, and by taking the ratio given in Equation (7). Physically, the lateral magnification is the ratio of the height of the image to the height of the object, both heights being measured perpendicularly to the optical axis. By defining lateral magnification by Equation (7), and remembering that y values of points below the optical axis have signs opposite to those above, we see that a posi-

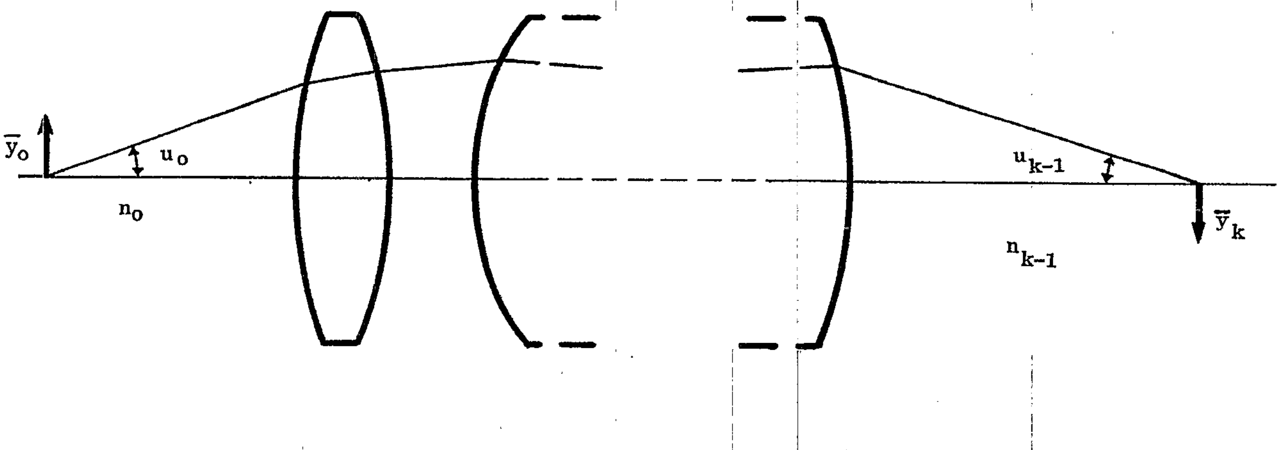


Figure 6.2 - Diagram illustrating the data used to compute the optical invariant.

tive value of m indicates an erect image. A negative value of m indicates an image inverted with respect to the object.

6.3.7 Angular magnification.

6.3.7.1 There are instruments which work with the object placed at a large distance t_0 from the first surface of the lens or mirror. If this distance is great enough to assume it is infinite, then the ray coordinates on the first surface for the axial and oblique rays are: y_1 ; $u_0 = 0$; $\bar{y}_1$; $\bar{u}_0$. The optical invariant, for the first surface (1) and the space to the left (0), becomes

$$\Phi = - y_1 \; (n_0 \; \bar{u}_0) \, .$$

In the image plane, $y_k = 0$, so

$$- y_1 \; (n_0 \; \bar{u}_0) = \bar{y}_k \; (n_{k-1} \; u_{k-1}) \, ,$$

and

$$\bar{y}_k = \frac{- y_1}{(n_{k-1} \; u_{k-1})} \; (n_0 \; \bar{u}_0) \, . \tag{8}$$

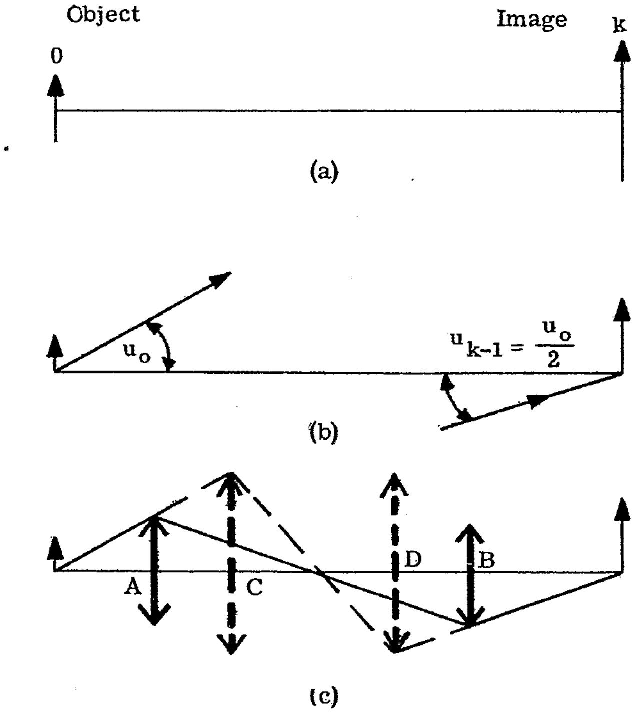
In visual instruments, the image surface is usually at a great distance from the last optical surface (k - 1). If the distance is assumed to be infinite, then $u_{k-1} = 0$, and

$$\Phi = - y_k \; (n_{k-1} \; \bar{u}_{k-1}) \, .$$

When both the object and image surfaces are assumed to be at infinity we have a telescopic system and the optical invariant is

$$\Phi = - y_1 \; (n_0 \; \bar{u}_0) = - y_k \; (n_{k-1} \; \bar{u}_{k-1}) \, .$$

The most familiar example of a telescopic system is a telescope for which both object and image surfaces are at infinity; when so adjusted the telescope is said to be afocal. From the material to be presented in a



(a)

(b)

(c)

Figure 6.3 - Diagrams illustrating the use of the Smith Helmholtz equations. Thin positive lenses are represented by the symbol ↕ , thin negative lenses by ⥯ .

later section we can say that such a telescope has its focal lengths equal to infinity and both focal points at infinity.

6.3.7.2 The angular magnification, α , is defined as the ratio $\bar{u}_{k-1}/\bar{u}_o$. Therefore the angular magnification for a telescope in afocal adjustment is

$$\alpha = \frac{y_1 \; n_o}{y_k \; n_{k-1}} = MP \; . \tag{9}$$

For a telescope, the angular magnification is called the magnifying power (MP).

6.3.8 The Smith-Helmholtz and the Lagrange equations. Equations (7) and (8) can be rewritten as

$$\bar{y}_o \; n_o \; u_o = \bar{y}_k \; n_{k-1} \; u_{k-1}$$

and

$$y_1 \; n_o \; \bar{u}_o = - \bar{y}_k \; n_{k-1} \; u_{k-1} \; .$$

These equations are referred to as the Smith-Helmholtz equations by some optical writers, and the LaGrange equations by others. Through the use of these equations, it is possible to decide rapidly what is needed to set up a given optical system. For example suppose we wish to form an erect image on surface k twice the size of the object on surface 0 . See Figure 6.3 (a) . Equation (7) shows that if m is to be $+ 2$ then u_o and u_{k-1} must have the same sign. This is illustrated in Figure 6.3 (b) for the case of $n_o = n_{k-1}$. A ray emerging from the base of the object at an angle u_o must pass through the optical system and emerge from below the optical axis at an angle $u_o/2$. As is shown in Figure 6.3 (c), this can be accomplished by any number of methods. A positive lens may be placed at A and be adjusted to refract the rays to cross the axis. At B a second positive lens refracts the rays to the final image. On the other hand two lenses could be used at C and D if desired, in which case the axial rays would refract as shown by the dotted lines.

6.4 LINEARITY OF THE PARAXIAL RAY TRACING EQUATIONS

6.4.1 General. In Sections 5.9.3 and 5.10 we have seen that finite heights and angles can be used with the paraxial ray trace equations. The basic reason for this is that these equations, 5-(56) and 5-(57), are linear. Another result of this linearity is that if two rays are traced through an optical system, it is possible to predict the path of any other paraxial ray. The proof of this fact will be developed below.

6.4.2 Proof of the theorem.

6.4.2.1 In order to prove the statements given above, let y and $\bar{y}$ be the heights of any two paraxial rays on the j th surface. Corresponding to these two rays, u and $\bar{u}$ are the angles between the rays and the optical axis. If $\bar{\bar{y}}$ and $\bar{\bar{u}}$ are the height and slope angle of any third ray, we wish to show that the equations

$$A \; \bar{y} + B \; y = \bar{\bar{y}} \tag{10a}$$

and

$$A \; \bar{u} + B \; u = \bar{\bar{u}} \tag{10b}$$

are valid for the entire optical system. We also must be able to calculate the values of A and B.

6.4.2.2 Equation (10a) applies to the jth surface. Using Equation 5-(56), we can show that an equation similar to (10a) applies to the $j + 1$ surface. Substituting Equation 5-(56) into Equation (10a) gives

$$\bar{\bar{y}}_{+1} - \frac{t}{n}(n\bar{\bar{u}}) = A \; \bar{y}_{+1} - A \frac{t}{n}(n\bar{u}) + B \; y_{+1} - B \frac{t}{n} \; nu \; .$$

Collecting the terms involving n results in the expression $t(\bar{\bar{u}} - A\,\bar{u} - B\,u)$. But this equals zero by Equation (10b) so that

$$\bar{\bar{y}}_{+1} = A \; \bar{y}_{+1} + B \; y_{+1} \; .$$

Hence Equation (10a) holds for the $j + 1$ surface, and therefore, by induction, for any and all surfaces.

6.4.2.3 Similarly, we show that Equation (10b) holds for all spaces. Substituting Equation 5-(57) into Equation (10b), and collecting terms, we have

$$\frac{n_{+1}}{n}\bar{\bar{u}}_{+1} = \frac{n_{+1}}{n} A \bar{u}_{+1} + \frac{n_{+1}}{n} B u_{+1} + \frac{c(n - n_{+1})}{n}(\bar{\bar{y}} - A\bar{y} - By).$$

By Equation (10a) the last term equals zero. Hence

$$\bar{\bar{u}}_{+1} = A\bar{u}_{+1} + B u_{+1},$$

and Equation (10b) applies to any and all spaces.

6.4.2.4 We have shown that Equations (10a) and (10b) apply to all surfaces and all spaces respectively and hence to the entire optical system. Solving these equations for A and B gives

$$A = \frac{\bar{\bar{y}} u - \bar{\bar{u}} y}{\bar{y} u - \bar{u} y} = n(\bar{\bar{y}} u - \bar{\bar{u}} y)/\Phi$$

and

$$B = \frac{\bar{y}\bar{\bar{u}} - \bar{u}\bar{\bar{y}}}{\bar{y} u - \bar{u} y} = n(\bar{y}\bar{\bar{u}} - \bar{u}\bar{\bar{y}})/\Phi.$$

These equations hold for any surface and the space to the right of that surface. In particular, we will use the expression for A for the object surface, and that for B for surface number 1.

6.4.3 Two particular rays.

6.4.3.1 Because the theorem proved in Section 6.4.2 holds for any three rays, we can choose these rays in

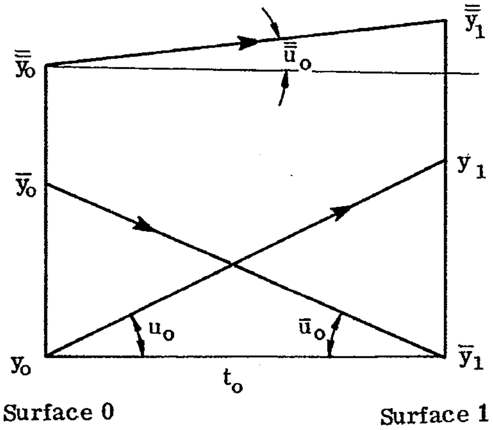


Figure 6.4 - Rays used to find simple expressions for A and B.

such a way as to simplify the calculation of A and B. The two particular rays we use are: (1) any ray from the center of the object surface ($x_o = y_o = 0$), and (2) any ray from the object ($\bar{y}_o \neq 0$) which intersects the axis at the center of the first surface ($\bar{x}_1 = \bar{y}_1 = 0$). These two particular rays, and any third ray, are shown in Figure 6.4. Using $y_o = \bar{y}_1 = 0$, the expressions for A and B reduce to

$$A = \bar{\bar{y}}_o / \bar{y}_o ,$$

and

$$B = \bar{\bar{y}}_1 / y_1 . \tag{11}$$

Using Figure 6.4, we have

$$y_1 = u_o t_o = - u_o \bar{y}_o / \bar{u}_o ,$$

and therefore

$$B = \frac{- \bar{\bar{y}}_1 \bar{u}_o}{\bar{y}_o u_o} . \tag{12}$$

6.4.3.2 The two particular rays chosen are often specified more stringently. In order to get some idea as to the necessary diameters of the elements, the ray from the axial object ($y_o = 0$) is taken at a value of u_o so as to pass through the edge of the aperture stop. Such a ray is called a rim ray, or marginal ray; the value of u_o determines the energy passing through the system. The other ray is taken as coming from the top of the object. This gives some idea as to the diameters of the elements necessary to attain the desired field of view. We will specify later that this second ray ($\bar{y}_k \neq 0$) be the chief ray.

6.4.3.3 The above two paragraphs have specified the two particular rays ($y_o = 0$ and $\bar{y}_1 = 0$) be chosen so as to easily evaluate A and B from the known data and the initial third ray data. (It should be emphasized that this is not necessary; any two rays and the initial third ray data will suffice to determine A and B). Instead of choosing particular values of y_o and $\bar{y}_1$, we could have chosen particular values of u_o and $\bar{u}_1$, for example 0 . This would result in $A = \bar{\bar{u}}_o / \bar{u}_o$ and $B = \bar{\bar{u}}_1 / u_1$. Note the correspondence between these and the equations in Paragraph 6.4.3.1.

6.5 THE CARDINAL POINTS OF AN OPTICAL SYSTEM

6.5.1 General.

6.5.1.1 We have already seen, in Sections 5.9.3, 5.10, and 6.4, some important consequences of the linearity of the paraxial ray trace equations. Another consequence, to be discussed in Section 6.5, is the presence of certain special points which exist in any optical system. Six of these points, all lying on the optical axis and known as the cardinal points, are of great usefulness in analyzing an optical system. The reason why the linearity of the paraxial ray equations lead to the existence of the cardinal points will not be developed in detail. It may be mentioned here, however, that the equations which we will develop from the concept of the cardinal points can be derived directly from the ray trace equations. One such equation, for example, was derived in Paragraph 6.2.3 . The fact that both the paraxial ray equations and the assumption of the existence of cardinal points lead to the same equations is indicative of the connection between Sections 6.4 and 6.5.

6.5.1.2 The cardinal points, and the letters used to designate them, are as follows:

(a) The first and second focal points, F_1 and F_2 .

(b) The first and second principal points, P_1 and P_2 .

(c) The first and second nodal points, N_1 and N_2 .

Sometimes the words first and second are replaced by primary and secondary, or by object and image, respectively.

6.5.2 The second focal point and the second focal length. In the sample calculation shown in Table 6.6, if the axial ray is traced from an infinitely distant object, $t_o = \infty$ and $u_o = 0$. This ray will pass through the optical system and eventually cross the axis at what is called F_2 , the second focal point. (See

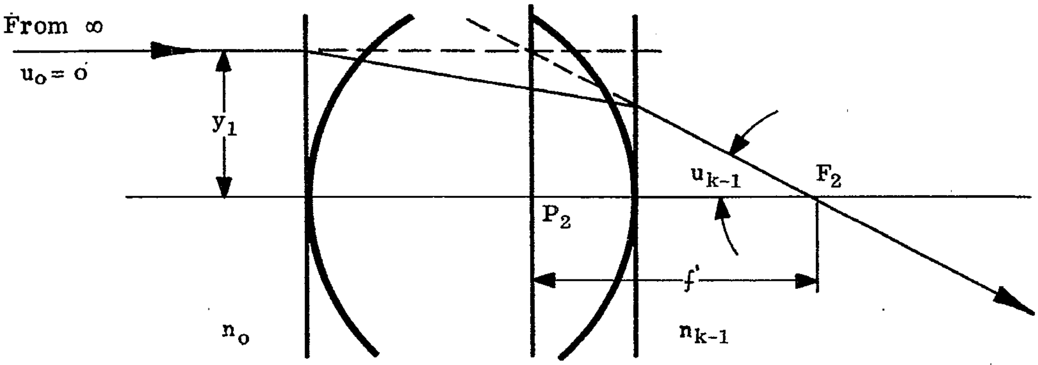


Figure 6.5 - Location of second focal point, second principal point, and second focal length.

Figure 6.5) . The second focal point is, therefore, the intersection (in image space) of the optical axis and a ray which (in object space) was initially parallel to the optical axis. This cardinal point can also be considered as the axial image of an infinitely distant axial object. This is why it is sometimes referred to as the image focal point. Because the height of the axial ray, y_1 , is arbitrary, all rays parallel to the optical axis, coming from an object surface, intersect at the second focal point. We can think of an image surface, intersecting the axis at F_2 . This is the second focal surface, which for paraxial rays becomes the second focal plane. Then $y_k = 0$, and Equation (8) applies,

$$\bar{y}_k = - y_1 \, n_o \, \bar{u}_o / (n_{k-1} \, u_{k-1}) .$$

The second focal length is defined as,

$$f' = - y_1 / u_{k-1} . \tag{13}$$

Physically, the second focal length is the distance between the second focal point and the second principal point, defined below. The reason a telescope in afocal adjustment (see Paragraph 6.3.7.1) has an infinite (second) focal length is that $u_{k-1} = 0$. Hence the final axial ray is parallel to the axis, and F_2 is at infinity.

6.5.3 The second principal point. The second principal point is located by erecting a plane perpendicular to the optical axis at the point of intersection of the forward-extended entering ray and the backward-extended exit ray. The intersection of this plane (the second principal plane) with the optical axis is the second principal point, P_2 . From Figure 6.5 it can be seen that

$$f' = P_2 \, F_2 .$$

If the second principal point is to the left of the second focal point, f' is positive; otherwise it is negative.

6.5.4 The second nodal point. The second nodal point, N_2 , is also an axial point, as are F_2 and P_2 . It is a point such that the distance

$$N_2 \, F_2 = (P_2 \, F_2) \frac{n_o}{n_{k-1}} .$$

With this expression Equation (8) can then be written

$$\bar{y}_k = f' \frac{n_o \bar{u}_o}{n_{k-1}} = P_2 F_2 \frac{n_o \bar{u}_o}{n_{k-1}} = N_2 F_2 \bar{u}_o .$$

If $n_o = n_{k-1}$, then $P_2 F_2$ and $N_2 F_2$ are equal and the principal point P_2 and the nodal point N_2 coincide.

6.5.5 The first focal, principal and nodal points.

6.5.5.1 With similar arguments one can find a first focal point, F_1 , such that rays entering the system from F_1 will emerge from the last surface traveling parallel to the axis. For such an object point, $y_o = 0$, and $u_{k-1} = 0$. Therefore from the optical invariant equation,

$$\bar{y}_o = - y_k \frac{n_{k-1} \bar{u}_{k-1}}{n_o u_o} .$$

The first focal length f is now defined as,

$$f = \frac{y_k}{u_o} = F_1 P_1 . \qquad (14)$$

Finally using $F_1 N_1 = (F_1 P_1) \frac{n_{k-1}}{n_o}$, we have

$$\bar{y}_o = - f \frac{n_{k-1} \bar{u}_{k-1}}{n_o} = - F_1 P_1 \frac{n_{k-1} \bar{u}_{k-1}}{n_o} = - F_1 N_1 \bar{u}_{k-1} .$$

6.5.5.2 The physical meanings of the first focal and principal points, and the first focal length, correspond to those discussed in Sections 6.5.2 and 6.5.3. The first focal point (see Figure 6.6) is the intersection of the optical axis and a ray which will be parallel to the axis when it leaves the system. It is also the axial object whose axial image is infinitely distant. All rays parallel to the optical axis after emerging from the

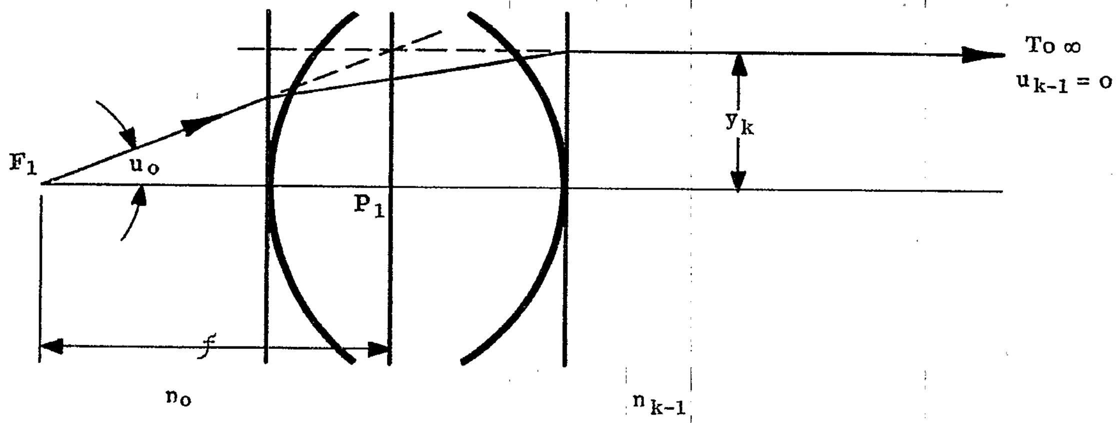


Figure 6.6 - Location of first focal point, first principal point, and first focal length.

system have passed through the first focal point. The plane perpendicular to the axis at F_1 is the first focal plane.

6.5.5.3 The first principal plane is a plane perpendicular to the optical axis passing through the intersection of the forward-extended ray through F_1 and the backward-extended ray emerging from the system parallel to the axis. The intersection of this plane with the axis is the first principal point. The first focal length is the distance between the first focal point and the first principal point, and is positive if F_1 lies to the left of P_1.

6.5.6 Object and image positions with respect to focal and principal points.

6.5.6.1 The previous sections, in connection with Figures 6.5 and 6.6, have explained the meaning of the focal and principal points, and the principal planes, from a graphical point of view. First, these ideas will be used to derive some well known relations between object and image positions. These relations will then be used to indicate additional characteristics of principal planes and nodal points.

6.5.6.2 Consider Figure 6.7 which indicates an object of height $\bar{y}_o$ at an arbitrary position. It should be emphasized here that Figure 6.7 indicates a general optical system, without reference to specific positions of refracting or reflecting surfaces. (Figures 6.5 and 6.6 show two refracting surfaces merely for concreteness; the ideas involved in those figures apply to the general system, as does the whole of Section 6.5). Of the infinite number of rays that come from the top of the object, we choose two whose course through the system we know from Figures 6.5 and 6.6. An entering ray, parallel to the optical axis, passes through F_2, and can be considered to be deviated only once, at the second principal plane. Similarly, a ray through F_1 exits parallel to the optical axis, and can be considered as having been deviated only once, at the first principal plane.

6.5.6.3 Four new distances are shown, Z, Z', S, and S'. Sign conventions are then established such that all these distances shown, as well as f and f', are positive. If any pair of points at the ends of the double arrows are reversed, the distance is negative. For example if the object is to the right of F_1, Z

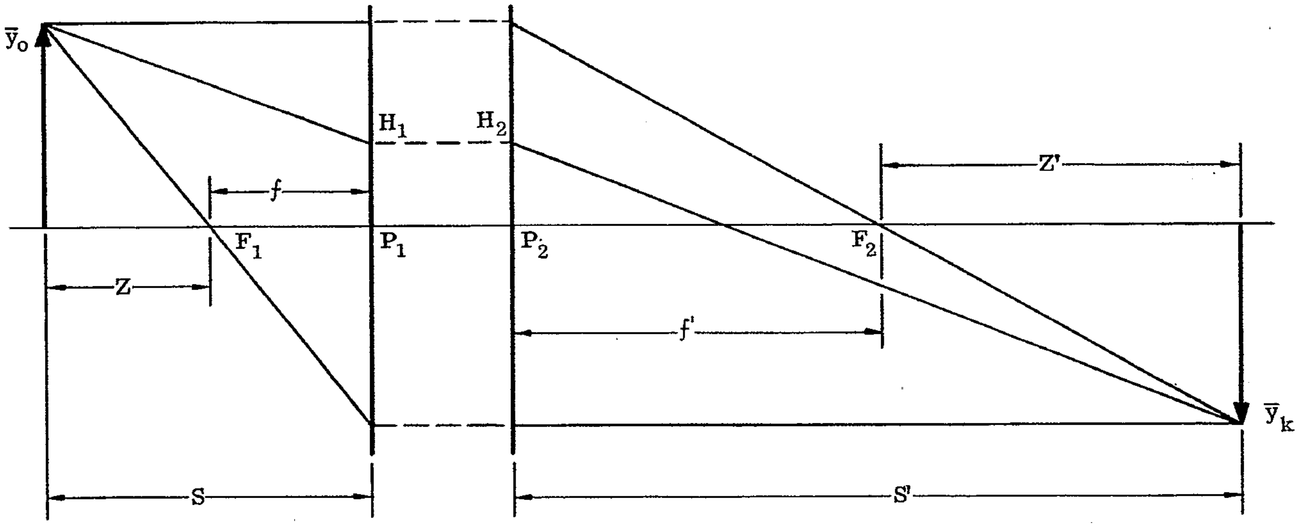


Figure 6.7 - Diagram showing object and image relations.

will be negative. From similar triangles, remembering that $\bar{y}_k$ is negative,

$$\frac{\bar{y}_k}{\bar{y}_o} = -\frac{Z'}{f'} = -\frac{f}{Z} .$$

Using the definition of lateral magnification, $m = \bar{y}_k / \bar{y}_o$, we have

$$m = -\frac{Z'}{f'} = -\frac{f}{Z} . \qquad (15)$$

Rearranging there follows

$$Z Z' = f f' . \qquad (16)$$

Equations (15) and (16) are in the Newtonian form, in which object and image positions are measured from the focal points, F_1 and F_2, respectively.

6.5.6.4 Another form of expressing these relations is the Gaussian form of these equations, in which object and image positions are measured from the principal points, P_1 and P_2, respectively. From Figure 6.7, $Z = S - f$ and $Z' = S' - f'$. Substituting these expressions into (15) and (16) gives

$$m = -\frac{S' - f'}{f'} = -\frac{f}{S - f}$$

and

$$(S - f)(S' - f') = f f' .$$

Expanding the last equation and dividing by SS', we have

$$\frac{f}{S} + \frac{f'}{S'} = 1 , \qquad (17)$$

and using (17), the lateral magnification becomes

$$m = -\frac{f S'}{f' S} . \qquad (18)$$

Equations (18) and (17) are in the Gaussian form and correspond to Equations (15) and (16). Whereas the latter pair does not involve S or S', and the former pair does not involve Z or Z', we may eliminate f and f' from Equation (16) by substituting $f = S - Z$ and $f' = S' - Z'$. The result is

$$\frac{Z}{S} + \frac{Z'}{S'} = 1 .$$

And using this with Equation (15), we have

$$m = -\frac{Z' S}{Z S'} .$$

6.5.6.5 It may be well to summarize here the specific meanings of the six distances used in the equations of Paragraphs 6.5.6.3 and 6.5.6.4. The sign conventions are included below if it is remembered that a distance measured to the right is positive.

f	is measured from F_1 to P_1 .
f'	is measured from P_2 to F_2 .
Z	is measured from the object plane to F_1 .
Z'	is measured from F_2 to the image plane.
S	is measured from the object plane to P_1 .
S'	is measured from P_2 to the image plane.

6.5.7 Additional characteristics of principal planes. Suppose the object is placed at the first principal plane. This means that $Z = -f$, and Equation (16) gives $Z' = -f'$. But this also means that the image is at the second principal plane; the two principal planes are therefore conjugate planes and P_1 and P_2 are conjugate points. (Equation (17) could have been used, with $S = 0$, giving $S' = 0$, which again locates the image at P_2). Using Equation (15) we find for this case $m = 1$. The two principal planes are therefore planes of unit positive magnification. This fact is very useful since it allows us to say that any point on the plane through P_1 is imaged at the same height on the plane through P_2. Therefore any other ray (see Figure 6.7), entering the system so that it intersects the first principal plane at H_1, exits from the system as if it came from H_2, at the same distance from the axis.

6.5.8 Additional characteristics of nodal points.

6.5.8.1 There is an important relation between the focal lengths of any optical system, and the refractive indices of object and image space. Equation (7) can be rewritten, using Figure 6.7, to give

$$m = \frac{n_o u_o}{n_{k-1} u_{k-1}} = \frac{n_o}{n_{k-1}} \left(-\frac{S'}{S}\right).$$

Comparing this with Equation (18) we have

$$f/n_o = f'/n_{k-1} . \qquad (19)$$

6.5.8.2 Equation (19) can be used to indicate a useful property of the nodal points. Using the expressions for $N_2 F_2$ and $F_1 N_1$ given in Sections 6.5.4 and 6.5.5, in connection with Figure 6.8, we have

$$P_1 N_1 = F_1 N_1 - F_1 P_1 = f\left(\frac{n_{k-1}}{n_o} - 1\right) = \frac{f}{n_o} (n_{k-1} - n_o),$$

and

$$P_2 N_2 = P_2 F_2 - N_2 F_2 = f'\left(1 - \frac{n_o}{n_{k-1}}\right) = \frac{f'}{n_{k-1}} (n_{k-1} - n_o).$$

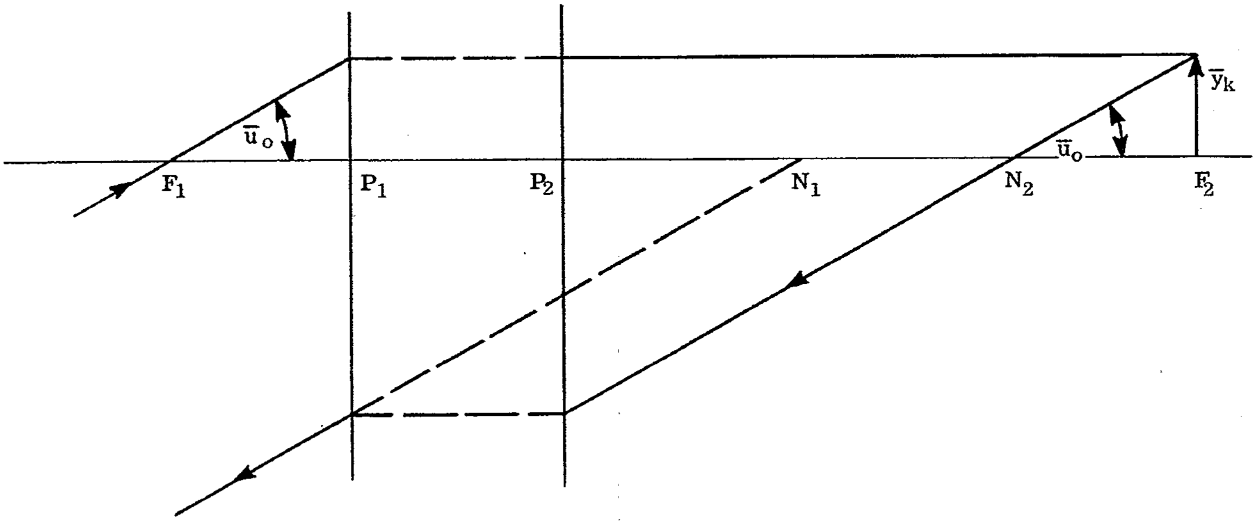


Figure 6.8 -Graphical construction to locate positions of nodal points.

Because of Equation (19), the following relations hold between the cardinal points.

$$P_1 N_1 = P_2 N_2 ,$$

$$P_1 P_2 = N_1 N_2 ,$$

$$F_1 N_1 = f' ,$$

$$N_2 F_2 = f .$$

And for object and image media the same, $n_o = n_{k-1}$, $f = f'$, and the principal and nodal points coincide, P_1 with N_1, and P_2 with N_2.

6.5.8.3 Because $P_1 P_2 = N_1 N_2$, two parallel lines, one through each nodal point, will intersect the principal planes in points equidistant from the axis. Hence these two rays are conjugate rays, and we have the important fact that any ray in object space which is heading toward N_1 will emerge from the system in the same direction from N_2. This gives us a graphical method for locating the nodal points, shown in Figure 6.8. A ray is shown entering the system at an angle $\bar{u}_o$ headed towards F_1 until it intersects the plane at P_1. It then emerges from the plane at P_2 parallel to the axis at the image height $\bar{y}_k$. A ray then traced backwards at an angle u_o with the axis must emerge anti-parallel to the entering ray as shown in the illustration, because all rays leaving a point on the focal plane are parallel to each other after emerging from the system. The two points N_1 and N_2 are the intersections with the axis of the two segments of this backwards traced ray.

6.5.9 Numerical example. A numerical example, represented in Figure 6.9, shows the location of the cardinal points of a lens with water on one side and air on the other. Given the three indices, two curvatures, and lens thickness, all other numerical values can be found using the equations already developed. An axial ray is traced through the system at $u_o = 0$ and y_1 arbitrary. t_2 can be found, using $y_3 = 0$. Therefore F_2 is located with respect to the second surface of the lens. A corresponding trace locates F_1. Equations (13) and (19) give f' and f respectively. The principal points and nodal points can now be located.

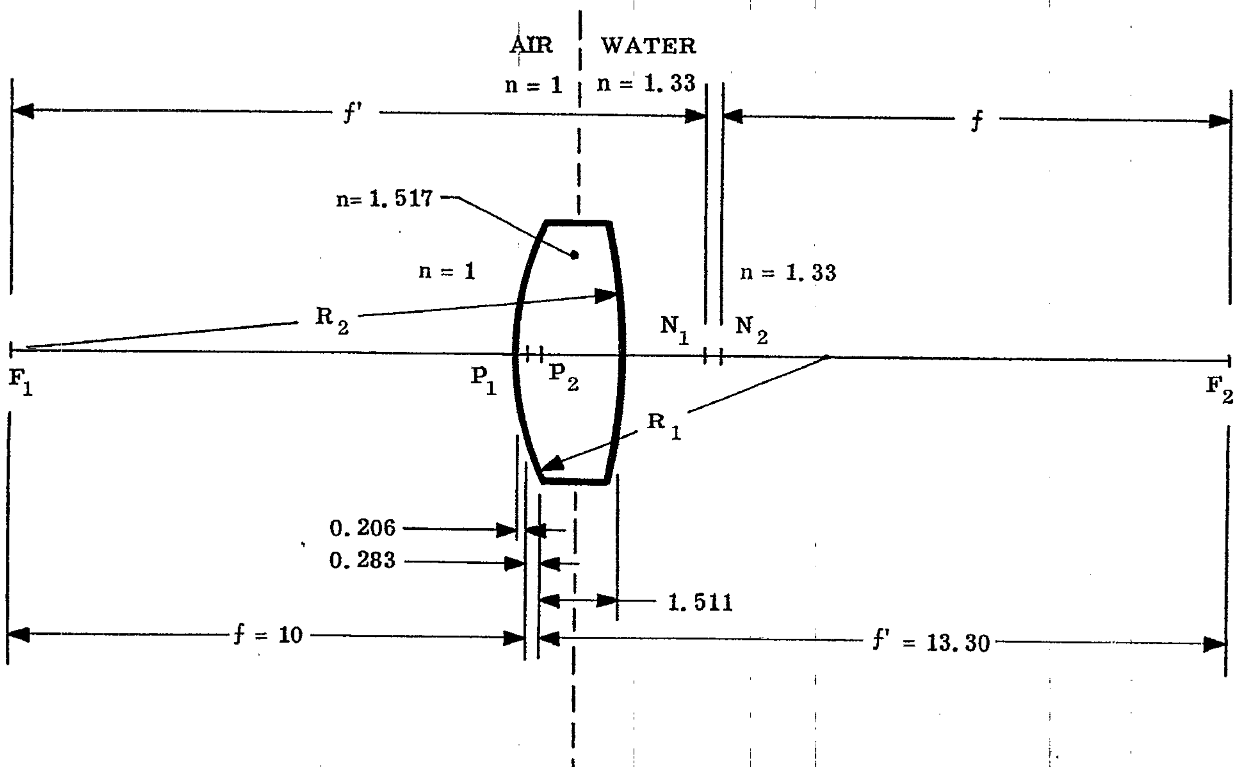


Figure 6.9. Numerical example showing location of the cardinal points for a lens with water on one side.

6.6 CALCULATION OF THE FOCAL LENGTH FROM FINITE CONJUGATE DATA

6.6.1 General. If an optical system is to be used at infinite conjugate, that is either the object or image or both are at infinity, then the entering axial ray is traced at $u_o = 0$, y_1 arbitrary. (For systems having the image at infinity for a finite object, the design is considered as if the rays went backwards through the system. Systems are therefore designed with the infinite conjugate as object, whether or not this agrees with the physical situation. The justification for this is that an optical system is reversible in the sense that rays traverse the same path in either direction). The ray trace automatically gives the focal length, f', by using Equation (13).

6.6.2 Finite conjugates. However, if the system images a finite conjugate object, and an axial ray and an oblique paraxial have been traced, Equation (13) does not apply. It is possible, nevertheless, from the data obtained from these two rays, to calculate the focal length. If two rays have been traced as shown in Figure 6.4 and in the presentation of Table 6.6, then

$$A = \bar{\bar{y}}_o / \bar{y}_o$$

and

$$B = -\bar{\bar{y}}_1 \ \bar{u}_o / \bar{y}_o \ u_o \ ,$$ this latter being Equation (12).

With these constants known, it is possible to predict the final $\bar{\bar{u}}_{k-1}$ for a ray entering the lens parallel to the axis. For then $\bar{\bar{u}}_o = 0$ and $\bar{\bar{y}}_o$ ($= \bar{\bar{y}}_1$) are the initial conditions for the third ray.

Now writing Equation (10b) for the final angle,

$$\bar{\bar{u}}_{k-1} = \frac{\bar{\bar{y}}_o}{\bar{y}_o} \bar{u}_{k-1} - \frac{\bar{\bar{y}}_1 \ \bar{u}_o}{\bar{y}_o \ u_o} u_{k-1} .$$

From this equation, and Equation (13) written for the third ray, we have

$$f' = - \frac{\Phi}{n_o (u_o \ \bar{u}_{k-1} - \bar{u}_o \ u_{k-1})} \tag{20}$$

where $\Phi = \bar{y}_o \ (n_o \ u_o)$ from Paragraph 6.3.6.

6.7 SYSTEMS OF THIN LENSES IN AIR

6.7.1 Concept of the thin lens.

6.7.1.1 None of the basic material presented so far presupposes any specific form of the optical system other than that it is a centered system. We now want to specialize the system somewhat and consider a single lens, an example of which is shown in Figure 6.9. In that example, $n_o \neq n_2$, so that $f \neq f'$. If $n_o = n_2 = 1$, the lens is in air, and $f = f'$. The nodal and principal points coincide as explained in Paragraph 6.5.8.2. Because of the equality of the two focal lengths, Equations (15) through (18) can be simplified.

6.7.1.2 An additional simplification can be attained by assuming that the axial lens thickness, t_1 in the above example, is small compared with t_o and t_2. If t_1 can be neglected, the lens is called a thin lens. Since the two deviations of the ray are considered to occur at one point, for a thin lens, both principal planes coincide with the lens of zero thickness. For this case, S and S' are the distances measured to the intersection of the lens with the optical axis, and Equations (17) and (18) take the familiar form for a thin lens in air. The two nodal points also coincide with the lens; hence a ray directed towards the lens center will emerge from the same point in the same direction. In some special cases, such as high curvature meniscus lenses (highly warped lenses), the thickness may be small, but not completely negligible. In these cases the lens may be "thin" for certain applications (for example, calculation of focal length), but not "thin" for others (for example, calculation of principal points positions). In such intermediate cases, where the lens is neither completely thick or completely thin, the principal and nodal points do not necessarily coincide with the center of the lens.

6.7.2 Focal length and power of a thin lens in air. Many optical systems are made up of individual two-surface lenses separated by air. Paraxial rays can, of course, be traced through any system of this type by using Equations 5-(56) and 5-(57), but considerable simplification can be made if it can be assumed that the individual lenses are thin. In the layout shown in Table 6.8, an axial paraxial ray and an oblique paraxial ray are traced through a thin, two-surface element in air. For the axial paraxial ray, we have

$$u_o = y/t_o \,,$$

$$u_2 = y/t_o + y(1-n)c_1 + y(n-1)c_2 \,,$$

and

$$u_2 = u_o - y(n-1)(c_1 - c_2) \,. \tag{21}$$

The focal length may be calculated from Equation (20), using $\Phi = \bar{y}_o\,(n_o u_o)$. If numerical calculations are made, the data are found in a table similar to Table 6.8. Therefore,

$$f' = -\frac{\Phi}{n_o\,(\bar{u}_2 u_o - \bar{u}_o u_2)} = \frac{t_o \bar{u}_o u_o}{(\bar{u}_o u_o - \bar{u}_o u_2)} \,,$$

or

$$f' = \frac{t_o u_o}{u_o - u_2} = \frac{1}{(n-1)(c_1 - c_2)}$$

and

$$1/f' = (n-1)(c_1 - c_2) = \phi \,. \tag{22}$$

Equation (22) is the well known formula for the focal length of a thin lens in air. It is more convenient to use it in the latter form, where ϕ is called the power of the thin lens.

SURFACE	Object		1		2		3
c	c_o		c_1		c_2		c_3
t		t_o		0		t_2	
n		1		n		1	
$(n_{-1}-n)c$	0		$(1-n)c_1$		$(n-1)c_2$		
t/n		t_o		0		t_2	
y	0		y		y		0
nu		y/t_o		$y(1-n)c_1$ $+ y/t_o$		$y(n-1)c_2$ $+y(1-n)c_1$ $+ y/t_o$	
$\bar{y}$	$-t_o\bar{u}_o$		0		0		
$n\bar{u}$		$\bar{u}_o$		$\bar{u}_1$		$\bar{u}_2$	

$\bar{u}_o = \bar{u}_1 = \bar{u}_2$

Table 6.8- Paraxial rays traced through a thin lens.

6.7.3 Ray trace equations for thin lens systems in air.

6.7.3.1 Equation (21) can be written

$$u_2 = u_o - y\phi \,.$$

The similarity between this and Equation 5-(57) is now apparent. Equation 5-(56) can be used to transfer between lenses. We have then the transfer and refraction equations for thin lens systems. These equations, (23) and (24), are written for a general thin lens j.

$$y = y_{-1} + t_{-1}\, u_{-1} \,, \tag{23}$$

$$u = u_{-1} + y(-\phi) \,. \tag{24}$$

Table 6.9 illustrates a method using Equations (23) and (24) for calculating the familiar expression for the focal length of a dialyte, i.e., two thin lenses separated by the distance d .

SURFACE	LENS (a)	LENS (b)	IMAGE
$-\phi$	$-\phi a$	$-\phi b$	
d		d	
y	y_1	$(1-d\phi a)y_1$	
u	0	$-\phi a\, y_1$	$(-\phi a-\phi b+d\phi a\phi b)y_1$

$$\frac{1}{f'} = \phi = -\frac{u_{k-1}}{y_1} = \phi a + \phi b - d\,\phi a\,\phi b .$$

Table 6.9 - Tracing a paraxial ray, $u_o = 0$ and y_1 arbitrary through two thin lenses.

6.7.3.2 The tracing of paraxial rays through thin lens systems is probably the one remaining calculation that lens designers do on desk calculators. In optical design work, a great deal of time and thought must necessarily go into the preliminary layout work. The designer must decide where to place the lenses, and what focal lengths are to be used. He needs to know approximately the sizes of lenses needed, and the approximate path of rays as they pass through the system. All these calculations can be made assuming thin lenses, and it is a problem so varied that it does not lend itself well to a large computer. Experience shows that desk calculators or slide rules are preferred at this stage of the design.

6.8 OPTICAL SYSTEMS INVOLVING MIRRORS

6.8.1 Sign conventions. It was pointed out in Section 2.3.3 that the equation of refraction could be used for reflection by merely writing

$$n_{+1} = -n .$$

If this is done in all the refraction equations, they can be used for reflection. If a mirror is inserted in an optical system, it reflects the ray backwards so that if the light was originally traveling from left to right, it will travel from right to left after reflection. It is possible to treat reflecting surfaces in exactly the same way as refraction surfaces by adopting the following rules:

(1) Write all the curvatures with the usual sign convention. If a single surface is encountered several times in a reflecting system, the radius is always considered to have the same sign.

(2) Whenever the light travels from right to left, insert the index and thickness with a negative sign.

6.8.2 A mirror system and its ray tracing format. A typical mirror and lens system is shown in Figure 6.10. The proper way to lay out the data for ray tracing is shown in Table 6.10. Actual rays as well as paraxial rays can then be traced through this system exactly as though it were only a refracting lens. If the light travels from right to left in the jth space one must remember that the index of refraction (n_j) is negative.

6.8.3 First order imagery in a mirror.

6.8.3.1 By using the above procedure it is now possible to readily work out the first order optics of a single mirror. The problem is illustrated in Figure 6.11, and worked out in the presentation shown in Table 6.11. From the table, it is apparent, by applying Equation 5-(56), that

$$y_2 = 0 = 1 + (-t_1)\left(\frac{1}{t_o} + 2c\right) .$$

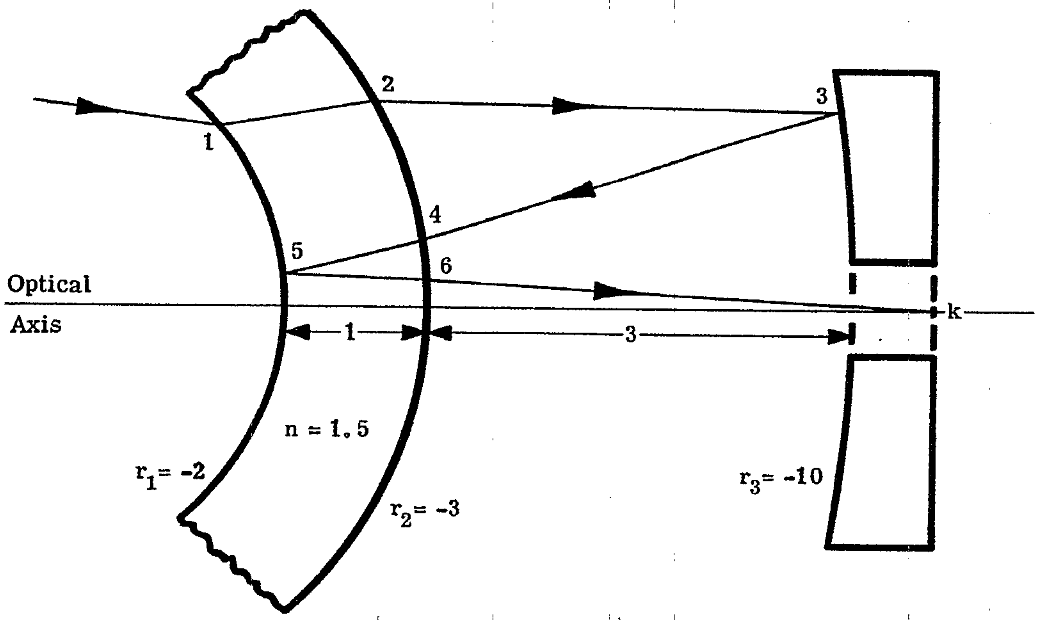


Figure 6.10 - The path of rays through a mirror system.

SURFACE	OBJECT		1		2		3		4		5		6		k
c	0		-0.500		-0.330		-0.100		-0.330		-0.500		-0.330		0
t		∞		1.000		3.000		-3.000		-1.000		1.000			
n		1.000		1.500		1.000		-1.000		-1.500		1.500		1.000	
$c(n_{-1} - n)$	0		0.250		-0.165		-0.200		-0.165		1.500		-0.165		0
t/n		∞		0.667		3.000		3.000		0.667		0.667			

Table 6.10 - Computing sheet format for mirror system illustrated above. Only the lens constants are included in the above table. The calculations, which are not given, are carried out as in Table 6.6.

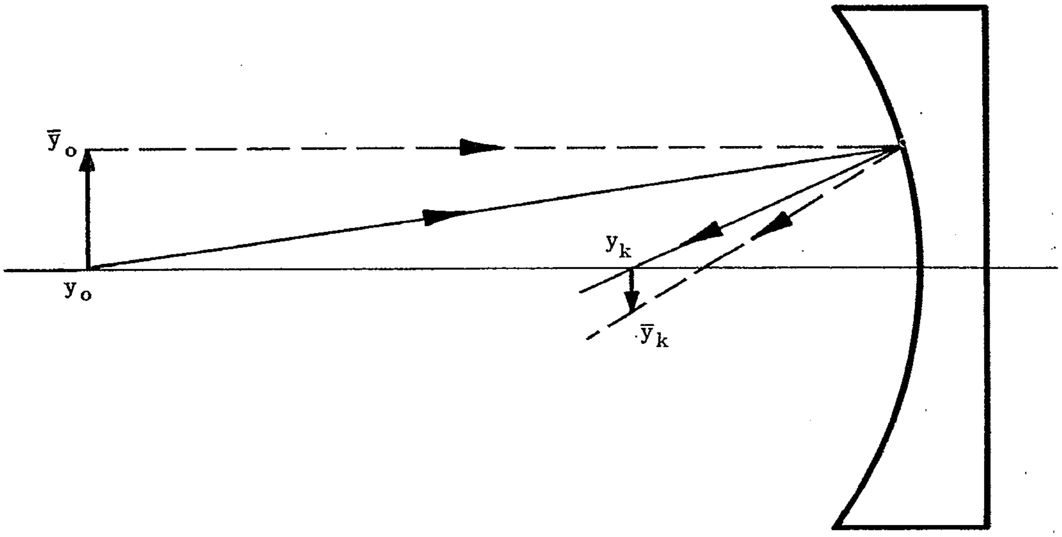


Figure 6.11 - Imaging an object in a concave mirror.

SURFACE	OBJECT		1		IMAGE
c	0		c		0
t		t_o		t_1	
n		1		-1	
$c(n_{-1} - n)$	0		2c		0
t/n		t_o		$-t_1$	
y	0		1		0
nu		$1/t_o$		$1/t_o + 2c$	
$\bar{y}$ *	1		1		$1 - t_1 2c$
$n\bar{u}$		0		2c	

* Ray traced parallel to axis to calculate focal length directly.

Table 6.11 - Ray tracing through a single mirror system.

Therefore

$$\frac{1}{t_1} = \frac{1}{t_o} + 2c = \frac{1}{t_o} + \frac{2}{r} . \qquad (25)$$

For a numerical example assume $r = -10$ and $t_o = 20$. Then $t_1 = -20/3$. The minus sign indicates that the image surface lies to the left of the mirror surface, as shown in Figure 6.11. The same equation could have been derived using Equation (1),

$$\frac{n_1}{t_1} + \frac{n_o}{t_o} = c_1 \left(n_1 - n_o \right)$$

and setting

$$n_1 = -n_o .$$

The magnification for the mirror may be found from Equation (7),

$$m = \frac{n_o \; u_o}{n_1 \; u_1} = \frac{1/t_o}{(1/t_o) + 2c} = t_1 / t_o .$$

The same equation could have been derived from Equation (18), remembering that $f' = -f$ because $n_1 = -n_o$.

6.8.3.2 The focal length of the mirror may be found by tracing a paraxial ray through the mirror at $\overline{y}_1 = 1$ and $\overline{u}_o = 0$ as noted in the lower two lines in Table 6.11. Equation (13) can be written

$$f' = -\frac{y_1 \; n_{k-1}}{(n_{k-1} \; u_{k-1})} ,$$

and used with the ray at $\overline{u}_o = 0$.

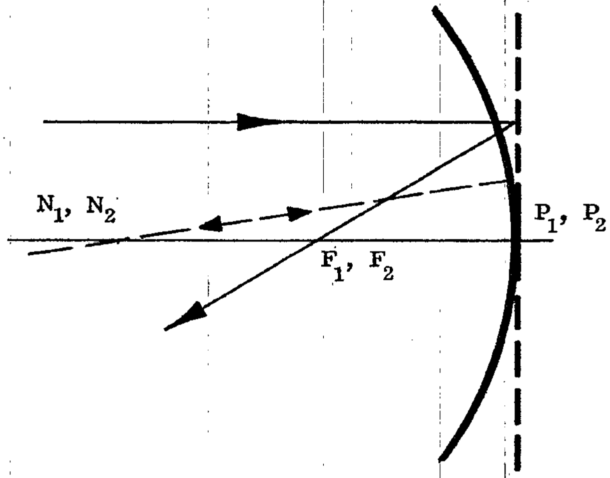


Figure 6.12 - The location of the principal points, focal points, and nodal points for a single mirror system.

Since $n_o = -n_{k-1}$, $\bar{y}_1 = 1$, and $(n_{k-1}\bar{u}_{k-1}) = 2c$, we have,

$$f' = \frac{n_o}{2c} = n_o \frac{r}{2} .$$

If r is negative as it is in Figure 6.11, f' is negative indicating that F_2 lies to the left of P_2. The equation $P_2 F_2 n_o = n_{k-1} N_2 F_2$ shows that for the mirror,

$$P_2 F_2 = -N_2 F_2 .$$

Since $P_2 F_2$ is negative for the example shown in Figure 6.11, then $N_2 F_2$ is positive. The location of P_1, N_1, F_1, P_2, N_2 and F_2 are shown in Figure 6.12. The nodal points are at the center of curvature.

6.9 DIFFERENTIAL CHANGES IN FIRST ORDER OPTICS

6.9.1 General.

6.9.1.1 The various steps followed in the design of an optical system are discussed in Section 9. The first two steps of the procedure are (1) selection of type of element for each part of the system, and (2) calculation of a first order thin lens solution. Step (2) involves the calculation of the focal lengths and separations of the individual elements, as well as first order aberrations which will be discussed in Section 6.10. The basic procedure for tracing paraxial rays, and therefore for determining focal lengths and spacings, have already been outlined in Section 6.

6.9.1.2 After the completion of step (2) the designer may feel that some changes are necessary so that the system meets more closely the required specifications. For example, he may have to change the focal length of the system. At the present stage of the system design (thin lens, paraxial rays), the designer can vary only the curvatures, the separations, and the indices of refraction. It therefore becomes important to know how changes in these three parameters affect the first order solution. In the remainder of Section 6.9 formulae will be given for computing the effects on first order optics for differential changes in the lens parameters.

6.9.2 Determination of the differential coefficients.

6.9.2.1 A change of any parameter, such as thickness, index of refraction, or curvature of a surface, will result in the paraxial ray changing its path to the next surface. Specifically, changes in t will change y_{+1}, and changes in n or c will change both u and y_{+1}. These changes will, in turn, cause changes on each surface up to and including the final image. The final changes, dy_k and du_{k-1}, which result from a change of any parameter associated with the jth surface, is certainly a function of changes dy_{j+1} and du_j. If the changes can be assumed to be differentials, it is possible to write

$$dy_k = \left(\frac{\partial y_k}{\partial y_{+1}}\right) dy_{+1} + \left(\frac{\partial y_k}{\partial u}\right) du \tag{26}$$

and

$$du_{k-1} = \left(\frac{\partial u_{k-1}}{\partial y_{+1}}\right) dy_{+1} + \left(\frac{\partial u_{k-1}}{\partial u}\right) du . \tag{27}$$

6.9.2.2 The partial derivatives in the above equations are called differential coefficients. If we trace two differential rays through the system, we have two values each for dy_{+1} and du (initial ray data) and two values each for dy_k and du_{k-1} (result of ray trace). Therefore, by tracing two differential rays near a given ray, it should be possible to determine the respective differential coefficients. It was shown in Section 5.9 that a paraxial ray is a differential ray traced near the optical axis. Therefore, we will use the axial paraxial ray and the oblique paraxial ray as the two differentially traced rays near the optical axis, taken as the given ray. It is possible then to evaluate the differential coefficients for changes in y_k and u_{k-1}, by making the following substitutions in Equations (26) and (27):

$$dy_k = y_k \qquad dy_{+1} = y_{+1} \qquad du = u \qquad du_{k-1} = u_{k-1}$$

$$d\bar{y}_k = \bar{y}_k \qquad d\bar{y}_{+1} = \bar{y}_{+1} \qquad d\bar{u} = \bar{u} \qquad d\bar{u}_{k-1} = \bar{u}_{k-1}$$

Two sets of simultaneous equations are thereby obtained. These equations, when solved for the derivatives, give:

$$\frac{\partial y_k}{\partial y_{+1}} = \frac{(\bar{y}_k \ u - y_k \ \bar{u})}{(\bar{y}_{+1} \ u - y_{+1} \ \bar{u})} = \frac{n(\bar{y}_k \ u - y_k \ \bar{u})}{\Phi} , \tag{28}$$

$$\frac{\partial y_k}{\partial u} = \frac{(y_k \ \bar{y}_{+1} - \bar{y}_k \ y_{+1})}{(\bar{y}_{+1} \ u - y_{+1} \ \bar{u})} = \frac{n(y_k \ \bar{y}_{+1} - \bar{y}_k \ y_{+1})}{\Phi} , \tag{29}$$

$$\frac{\partial u_{k-1}}{\partial y_{+1}} = \frac{(\bar{u}_{k-1} \ u - u_{k-1} \ \bar{u})}{(\bar{y}_{+1} \ u - y_{+1} \ \bar{u})} = \frac{n(\bar{u}_{k-1} \ u - u_{k-1} \ \bar{u})}{\Phi} , \tag{30}$$

and

$$\frac{\partial u_{k-1}}{\partial u} = \frac{(\bar{y}_{+1} \ u_{k-1} - y_{+1} \ \bar{u}_{k-1})}{(\bar{y}_{+1} \ u - y_{+1} \ \bar{u})} = \frac{n(\bar{y}_{+1} \ u_{k-1} - y_{+1} \ \bar{u}_{k-1})}{\Phi} . \tag{31}$$

6.9.3 Effect of curvature change on focal length.

6.9.3.1 The change in focal length, df', due to changes in curvature, thickness, and index is given by

$$df' = \left(\frac{\partial f'}{\partial c}\right) dc + \left(\frac{\partial f'}{\partial t}\right) dt + \left(\frac{\partial f'}{\partial n}\right) dn .$$

If the differential coefficients are known, then df' can be found for any small change in the system parameters. It will now be assumed that t and n are held constant.

6.9.3.2 Combining the transfer equation

$$y_{+1} = y + tu ,$$

with the above substitutions we have, for the case of $t = \text{constant}$,

$$dy_{+1} = t\,du .$$

Using this and Equations (28) to (31), Equations (26) and (27) become

$$dy_k = \frac{n}{\Phi} \left(y_k \ \bar{y} - \bar{y}_k \ y \right) du , \tag{32}$$

and

$$du_{k-1} = \frac{n}{\Phi} \left(\bar{y} u_{k-1} - y \bar{u}_{k-1} \right) du . \tag{33}$$

6.9.3.3 Equation (13), defining the second focal length, assumes that the axial paraxial ray was traced at $u_o = 0$. Differentiating this equation, remembering that y_1 is arbitrary and hence independent of c, we have

$$\frac{df'}{dc} = -\left(\frac{f'}{u_{k-1}}\right) \left(\frac{du_{k-1}}{dc}\right) = -\left(\frac{f'}{u_{k-1}}\right) \left(\frac{du_{k-1}}{du}\right) \frac{du}{dc} .$$

Differentiating 5-(57) it follows that

$$\frac{du}{dc} = \frac{y(n_{-1} - n)}{n} .$$

Therefore, using Equation (33),

$$\frac{df'}{dc} = \frac{\left[-f' (\bar{y} u_{k-1} - y \bar{u}_{k-1}) \right] \left[y (n_{-1} - n) \right]}{\Phi \, u_{k-1}} .$$

6.9.4 Effect of curvature change on final angle. In Table 6.12 a calculation is shown for a change in curvature made on the fourth surface of the example given in Table 6.6. Comparing the new u_6 with the original one in Table 6.6 we have $\Delta u_6 = 0.00469$. Now we will compare this value with a calculated value using the equations for the differential coefficients. Since we are making a change in the curvature only, keeping the thickness and index constant, we calculate

$$\frac{du_{k-1}}{dc} = \frac{du_{k-1}}{du} \frac{du}{dc} .$$

From Equation (32), and data from Table 6.6, the following calculation may be made,

$$\frac{du_6}{dc_4} = \frac{y_4 (n_3 - n_4)}{\Phi} (\bar{y}_4 u_6 - y_4 \bar{u}_6)$$

$$= - \frac{1.026 \times .621}{.5} (-0.00069 \times 0.08664 - 1.02606 \times 0.35886)$$

$$= 0.469 .$$

We have then that

$$\Delta u_6 = \frac{du_6}{dc_4} \Delta c_4 = (0.469)(0.01) = 0.00469 .$$

This is in exact agreement with the result from Table 6.12.

6.9.5 Effect of thickness change on final angle. It is also possible to compute the change in the final angle from a change in any thickness t. If t is changed, then,

$$dy_{+1} = u dt .$$

SURFACE		4		5		6		7
c		0.26973		0.05065		-0.24588		0
t			1.13691		0.6		14.9709	
n	1.621		1		1.620		1	
$c(n_{-1} - n)$		0.16750		-0.03140		-0.15245		0
t/n			1.13691		0.37037		14.9709	
y		1.02606		1.18794		1.22686		0
nu	-0.02948		0.14238		0.10508		-0.08195	

$\Delta u_6 = -0.08195 - (-0.08664) = 0.00469$

Table 6.12 - Calculations showing the effect on u_{k-1} of a change of $\Delta c_4 = 0.01$ in the data in Table 6.6

Therefore, using Equation (27) with $du = 0$, and Equation (30),

$$\frac{du_{k-1}}{dt} = \frac{\partial u_{k-1}}{\partial y_{+1}} \frac{dy_{+1}}{dt} ,$$

and

$$\frac{du_{k-1}}{dt} = \frac{nu}{\Phi} \left[\bar{u}_{k-1} \, u - u_{k-1} \, \bar{u} \right] .$$

6.10 CHROMATIC ABERRATION

6.10.1 The meaning of chromatic aberration. The variation of refractive indices with wavelength was discussed under the topic of dispersion in Section 2.6. The method of differential coefficients described in Section 6.9 can be used to calculate the effect of such a change in the index of refraction of the lenses. This change in index affects the refraction of each ray so that rays of different wavelengths pass through the system in slightly different paths. Generally these rays of different wavelengths give rise to more than a single image, a phenomenon called chromatic aberration. If the images are at different positions along the optical axis, the system exhibits longitudinal or axial chromatic aberration. If the images are of different lateral magnification, the system exhibits transverse or lateral chromatic aberration. Axial and lateral chromatic aberrations are sometimes referred to as axial color and lateral color, respectively.

6.10.2 Surface contributions.

6.10.2.1 As mentioned above, each surface introduces a certain amount of chromatic aberration appearing in the final image. The amount due to a particular surface is called the surface contribution. The general approach used to calculate first and third order aberrations is (1) determine the surface contribution, and (2) sum the contributions for all surfaces to find the total aberration. The individual contributions may be positive, negative, or zero. Hence the sum may be either positive, negative, or zero. In the last case the system would be free of this particular aberration.

6.10.2.2 The first order chromatic aberration contribution of any surface may be found by differentiating Equation 5-(57), assuming that $du_{-1} = 0$. This assumption means that the ray between the $j - 1$ and jth surfaces is unaberrated; hence we are considering only the contribution of the jth surface. The assumption $du_{-1} = 0$ also leads to $dy = 0$, because the ray to the left of the jth surface retains its original path. We then have,

$$n du + u dn = u_{-1} \, dn_{-1} + yc(dn_{-1} - dn) .$$

This can be put into a form more suitable for calculation. From Equation 5-(35), written for small angles, and Equation 6-(4), we have

$$i' = yc + u . \qquad (33a)$$

Using this equation, Equation 2-(1) for small angles, and Equation 6-(4), it is possible to derive the expression

$$du = i \frac{n_{-1}}{n} \left[\left(\frac{dn_{-1}}{n_{-1}} \right) - \left(\frac{dn}{n} \right) \right] .$$

6.10.2.3 Here, dn and dn_{-1} represent infinitesimal changes in index due to an infinitesimal change in wavelength λ. The change in u, due to a change of dn_{-1} and dn, will thus cause the ray to take a deviated path to the image. The change dy_k, in the final image, may then be calculated from

Equation (32). Since $y_k = 0$ for the axial ray, the value of dy_k is:

$$dy_k = - \frac{y \bar{y}_k n_{-1} i}{\Phi} \left[\left(\frac{dn_{-1}}{n_{-1}} \right) - \left(\frac{dn}{n} \right) \right] ;$$

$$dy_k = - \frac{y n_{-1} i}{(n_{k-1} u_{k-1})} \left[\left(\frac{dn_{-1}}{n_{-1}} \right) - \left(\frac{dn}{n} \right) \right] ;$$

$$dy_k = y n_{-1} i \left[\Delta \frac{dn}{n} \right] \Big/ (n_{k-1} u_{k-1}) ; * \tag{34}$$

or

$$dy_k = - a / (n_{k-1} u_{k-1}) .$$

The above derivation could have been equally well carried out for the oblique paraxial ray giving

$$d\bar{y}_k = y n_{-1} \bar{i} \left[\Delta \frac{dn}{n} \right] \Big/ (n_{k-1} u_{k-1}) = - b/(n_{k-1} u_{k-1}) \tag{35}$$

$$= dy_k \bar{i} / i . \tag{36}$$

6.10.3 Total chromatic aberration. Equation (34) gives the amount by which the image of an axial object point is displaced from the optical axis due to the jth surface. Similarly Equation (35) applies to the image of an object point off the axis. Both these equations give the transverse displacement in the final paraxial image plane due to changes dn_{-1} and dn. Now, if these changes are due to a change of wavelength $d\lambda$, changes dn and dn_{-1} occur at every surface in the lens. Each surface then contributes a dy_k and a $d\bar{y}_k$, and since they are all differentials, they are directly additive. The totals are

$$\text{total } dy_k = \text{TAch} = \frac{-1}{(n_{k-1} u_{k-1})} \sum_{j=1}^{j=k-1} a , \tag{37}$$

and

$$\text{total } d\bar{y}_k = \text{Tch} = \frac{-1}{(n_{k-1} u_{k-1})} \sum_{j=1}^{j=k-1} b , \tag{38}$$

where a and b are the chromatic surface coefficients. Note that Equation (34) has i while Equation (35) has $\bar{i}$. In all other terms the equations are identical. The symbols TAch and Tch have replaced dy_k and $d\bar{y}_k$ as descriptive terms to indicate the total transverse chromatic effects. TAch is the abbreviation for transverse axial chromatic aberration. Tch is the abbreviation for transverse chromatic aberration. A sample calculation for TAch is included in Table 6.7.

6.10.4 Particular wavelengths used to calculate chromatic aberration.

6.10.4.1 The first order chromatic aberration, strictly speaking, is the infinitesimal change, dy_k, resulting from a change dn which is due to a change $d\lambda$. Therefore, in order to calculate the infinitesimals, TAch and Tch, it is necessary to know the index at all wavelengths. As was discussed in Section 2.6.3, indices are measured at only certain standard wavelengths. It is possible to interpolate between standard wavelengths, using an appropriate dispersion formula, in order to calculate the index, and hence the chromatic aberration, at any wavelength.

6.10.4.2 However, in order to obtain accurate indices for ray tracing, it is customary to use only measured indices. Therefore in order to calculate dn, which is now considered a finite change, two wavelengths are chosen n_v and n_r. Then $dn_{v-r} = n_v - n_r$. [v and r indicate wavelengths at the ends (violet and red) of the visible region]. Then a wavelength λ_g between v and r is used as the reference index of refraction. λ_g is any wavelength in the middle part of the spectrum. The paraxial

---

*$\Delta(dn/n)$ is defined as $\left(\frac{dn}{n}\right) - \left(\frac{dn_{-1}}{n_{-1}}\right)$. The use of Δ is often used in optics to denote the difference between a quantity on the two sides of a refracting surface. For example, $\Delta n = (n - n_{-1})$.

rays are traced at wavelength λ_g . Therefore,

$$\mathrm{TAch}_{v-r} = (y_k)_v - (y_k)_r ,$$

and

$$\mathrm{Tch}_{v-r} = (\bar{y}_k)_v - (\bar{y}_k)_r .$$

The differences are measured in the paraxial image plane where $(y_k)_g = 0$. It should be pointed out that TAch_{v-r} and Tch_{v-r} tell only the difference in y_k and $\bar{y}_k$ for light at wavelengths λ_v and λ_r . In order to calculate other chromatic aberrations, for example $(y_k)_v - (y_k)_g$, the calculations are made with

$$dn_{v-g} = (n_v - n_g) .$$

The wavelengths chosen for calculation, depend on the wavelength region of interest. Visual optical systems are usually calculated with

$$n_v = n_F ,$$

$$n_g = n_D ,$$

and

$$n_r = n_C .$$

6.10.5 Graphical interpretation of axial and lateral color.

6.10.5.1 In Figure 6.13 a simple lens is shown with an exaggerated amount of chromatic aberration. A simple converging lens, which is necessarily uncorrected for aberrations, is said to be undercorrected. When a particular aberration is made zero, or smaller than some predetermined tolerance, the lens system is said to be corrected. If the aberration of the system has a sign opposite to that of a simple converging

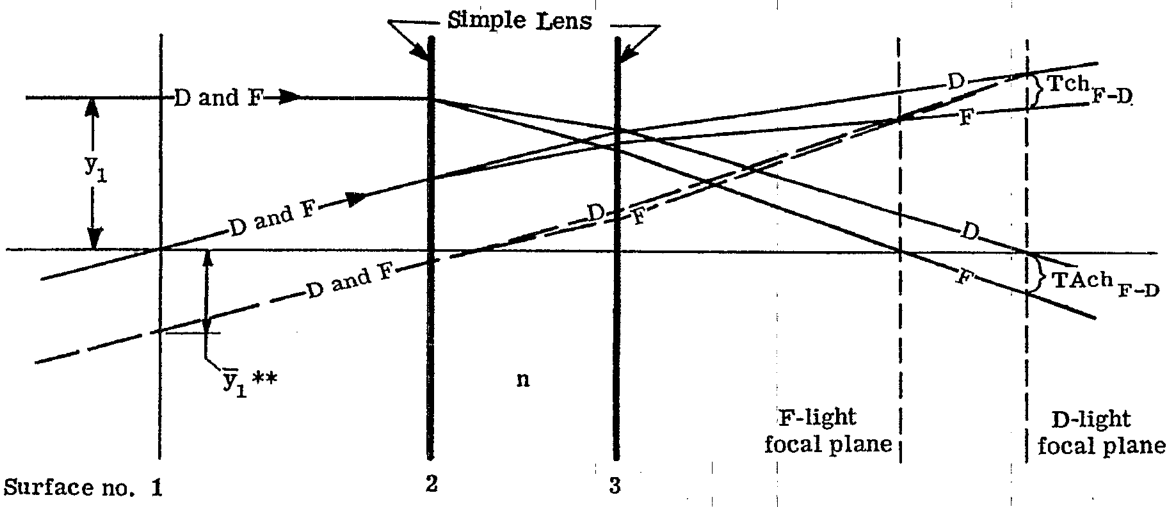


Figure 6.13 - Under-corrected chromatic abberation of axial and oblique rays in a simple lens.

lens, the system is over-corrected. The two surfaces of the lens in Figure 6.13 are labelled 2 and 3, and they appear as planes, as they should in the paraxial region. Axial and oblique rays in D and F light are shown as they pass through the lens. The oblique rays cross the axis at a reference surface #1. This reference plane will often coincide with the entrance pupil of the system. The pupils will be discussed in Section 6.11. With a positive lens, the F light image plane falls closer to the lens than the D light image plane. The chromatic blur, dy_{F-D} , is a linear function of y_1 , the height of the axial ray entering the system. This can be seen by considering Figure 6.13. All axial paraxial rays in D light pass through the same point on the optical axis, independent of y_1 . Hence all values of y_k for D light are zero and therefore Figure 6.14 indicates a horizontal line for D light. Similarly all axial paraxial rays in F light pass through a common point on the optical axis, independent of y_1 . Hence the separation of the two focal planes for F light and D light is a constant, independent of y_1 . This separation is called the longitudinal axial chromatic aberration, and is denoted by LA_{F-D} . From Figure 6.13 it is seen that

$$TAch_{F-D} = (LA_{F-D})\, u_{k-1} = -(LA_{F-D})\, \frac{y_1}{f'} .$$

Because the chromatic blur, $TAch_{F-D}$, is a linear function of y_1 , the line for F light in Figure 6.14 is straight and inclined to that for D light at the angle $(LA_{F-D})/f'$. Figure 6.14 shows a plot for $(y_k)_F$ and $(y_k)_D$ in the D light image plane, as a function of the height of the axial ray on the entrance pupil plane. This is a recommended way to indicate the transverse axial chromatic aberration of a system.

6.10.5.2 Figure 6.15 shows a plot of $\bar{y}_k$ versus $\bar{y}_1$ for F and D light. The chromatic blur, $d\bar{y}_{F-D}$, is a linear function of $\bar{y}_1$ for a reason similar to that given in Paragraph 6.10.5.1. For all values of $\bar{y}_1$, all D rays pass through a common point on the D light focal plane. Similarly, all F rays pass through a common point. Since the rays are paraxial, the oblique ray at $\bar{y}_1 = 0$ can be considered as an auxiliary axis; hence a ray parallel to it through a point $\bar{y}_1 \neq 0$ will make the same angle with the chief ray that an axial paraxial ray makes with the optical axis. The former angle is a linear function of $\bar{y}_1$, as u_{k-1} is a linear function of y_1 . Hence the chromatic blur is a linear function of $\bar{y}_1$, and the F light line is straight in Figure 6.15. The distance between the F and D chief rays in the D light image plane is as indicated in Figure 6.15. This is the value computed from Equation (38). The differ-

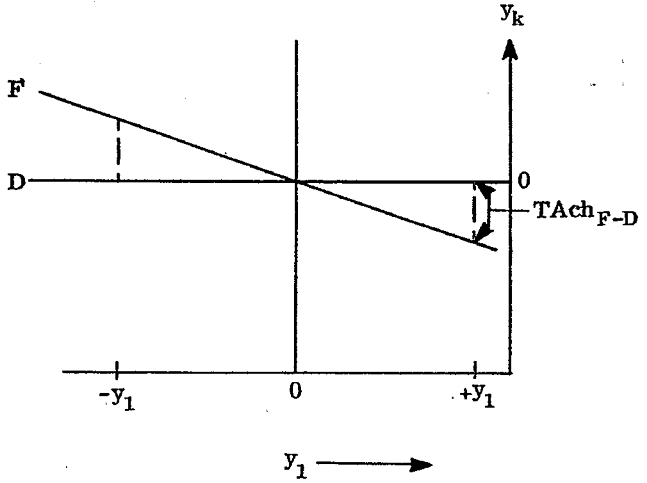


Figure 6.14 - A plot of y_k for F and D light versus the height y_1 of the axial paraxial rays on the entrance pupil plane.

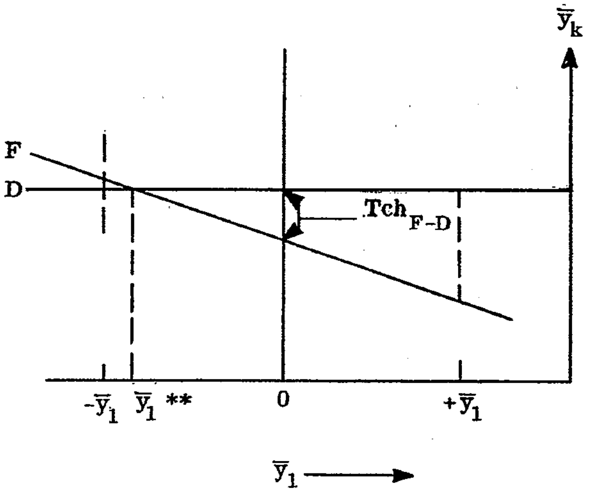


Figure 6.15 - A plot of $\bar{y}_k$ for F and D light versus the height $\bar{y}_1$ of the oblique paraxial rays.

ence in slope between the F and D lines is the same as for the axial rays shown in Figure 6.14, because the proportionality constant between TAch and y_1 is identical to that between Tch and $\bar{y}_1$. Figure 6.15 (and also Figure 6.13) shows that there is a value of $\bar{y}_1^{**}$ such that $\text{Tch}_{F-D} = 0$. This means that if the oblique paraxial ray had been taken through the lens at a value of $\bar{y}_1 = \bar{y}_1^{**}$, instead of $\bar{y}_1 = 0$, then Tch_{F-D} would come out to be zero. In fact, in general, it can be said that

$$\text{Tch}^*_{F-D} = \text{Tch}_{F-D} + \frac{\bar{y}_1^*}{y_1}\,\text{TAch}\,.$$

6.10.5.3 This equation states that Tch_{F-D} can be calculated for any oblique ray striking the entrance pupil plane at $\bar{y}_1^*$ with the above equation. The (*) is used to indicate the Tch for some oblique ray displaced from the ray passing through $\bar{y}_1 = 0$. Defining $\bar{y}_1^* / y_1 = Q$, the above equation may be written

$$\text{Tch}^*_{F-D} = \text{Tch}_{F-D} + Q\ \text{TAch}\,. \qquad (39)$$

Again it can be seen that it is necessary to trace only two paraxial rays through a lens system. It is possible to compute TAch, and Tch for any other rays from the data on these two.

6.10.6 Basic concepts in correcting systems for chromatic aberrations.

6.10.6.1 If two wavelengths, F and D for example, come to focus in the same image plane, then $(y_k)_F = (y_k)_D = 0$. This equation gives the condition for correction of the axial color. However, this does not mean that $(u_{k-1})_F$ will necessarily be equal to $(u_{k-1})_D$. If these two angles are not equal, then the magnifications between the object and image will not be equal, and $(\bar{y}_k)_F \neq (\bar{y}_k)_D$. Therefore, the system will have residual lateral color. Hence if both axial and lateral color are to be corrected, the rays in F and D light should emerge from the system at the same value of y_{k-1} and u_{k-1}.

6.10.6.2 The usual achromatic doublet lens is corrected for axial and lateral color because the axial rays in the F and D light never become significantly separated. See Figure 6.16 (a). In the case of two separated lenses, Figure 6.16 (b), it is clear that both elements must be color corrected, to keep the rays together all the way to the final image. If any axial color is allowed in the front element the rear element would have to be thick enough and designed properly to get the two rays together again before emerging from the rear surface. It is possible, by using the proper lens power and glass dispersion, to correct for axial and lateral color in widely spaced lenses as shown in Figure 6.16 (c). This is the principle used in the design of the famous Taylor triplet photographic lens. As a general principle, however, it is always advisable to keep the color rays as close together as possible at all times. This means, if the system is to be made up of several components, each component should be made achromatic.

6.10.7 Chromatic aberration in a thin lens.

6.10.7.1 It is possible to apply Equations (37) and (38) to a thin lens immersed in a non-dispersive medium and simplify the equations because the values of y and of $\bar{y}$ are the same on both surfaces. Suppose there is a thin lens in a system of thin lenses in air with values of y and $\bar{y}$ for heights of the axial and oblique paraxial rays. (See Figure 6.17). This lens will contribute the following amounts of axial and lateral color to the final image.

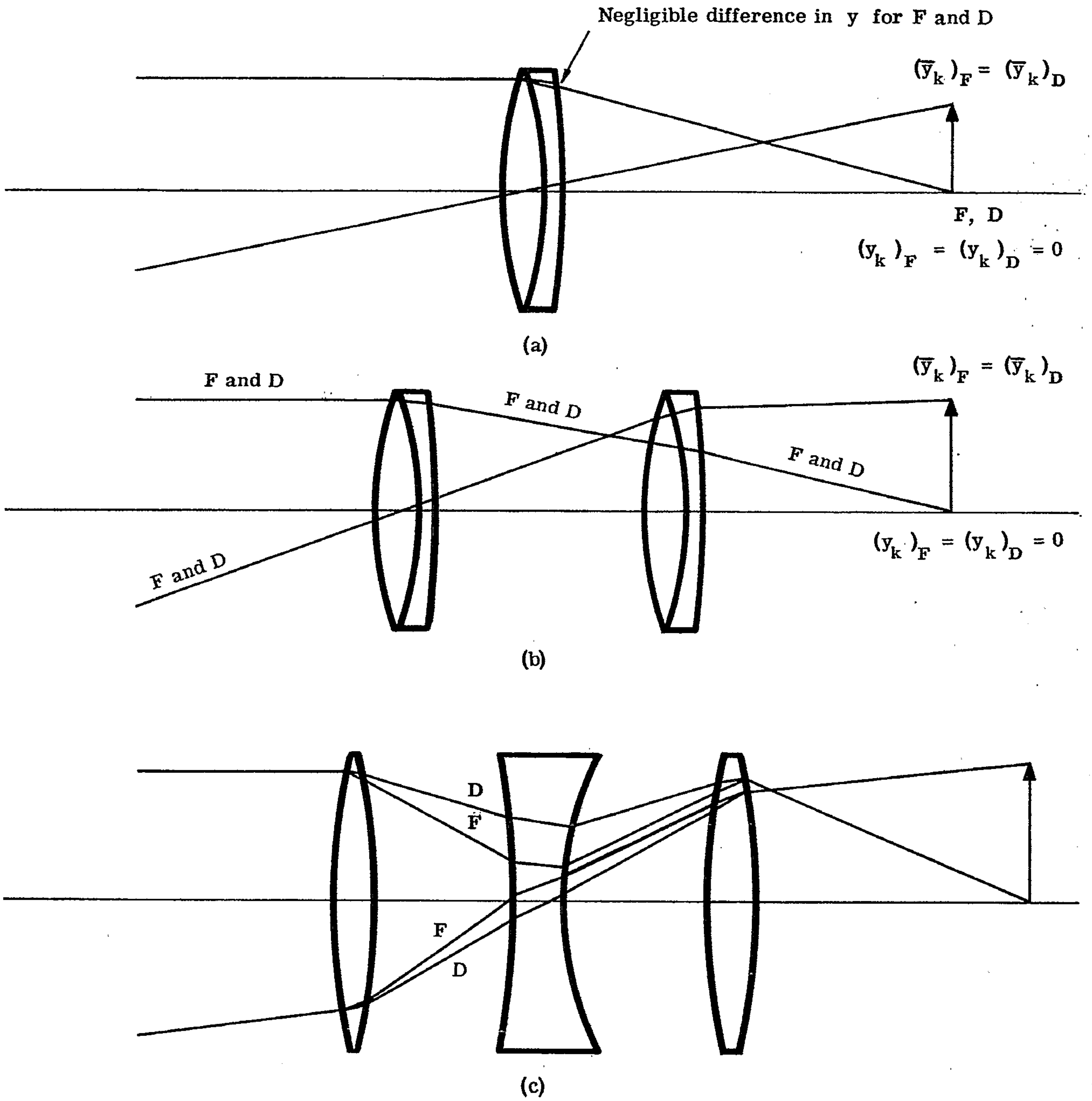
$$\text{TAch}_{v-r} = \frac{1}{(n_{k-1}\ u_{k-1})}\left(y^2\ \frac{\phi}{\nu_{v-r}}\right),$$

and

$$\text{Tch}_{v-r} = \frac{1}{(n_{k-1}\ u_{k-1})}\left(y\,\bar{y}\ \frac{\phi}{\nu_{v-r}}\right).$$

where ϕ is the power of the lens, and $\nu_{v-r} = (n_g - 1)/(n_v - n_r)$. These equations follow from Equations (37) and (38), with the use of Equations (4), (22), (33a) and 2-(1) for small angles.

6.10.7.2 Each of the thin lenses adds a contribution, so the final axial and lateral color for a system of η


Negligible difference in y for F and D
$(\bar{y}_k)_F = (\bar{y}_k)_D$
F, D
$(y_k)_F = (y_k)_D = 0$
F and D
F and D
F and D
F and D
$(\bar{y}_k)_F = (\bar{y}_k)_D$
$(y_k)_F = (y_k)_D = 0$
D
F
F
D

(a)

(b)

(c)

Figure 6.16 - Illustration of axial and lateral color correction for paraxial rays.

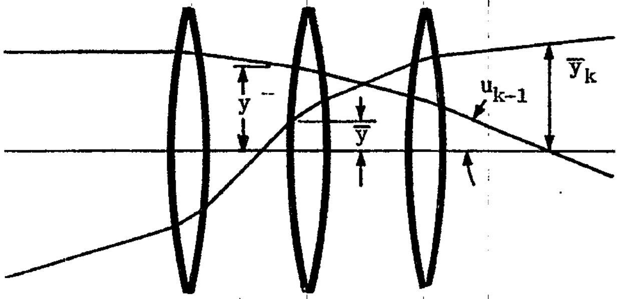


Figure 6.17 -A system of thin lenses.

SURFACE	(1, 2)	(3, 4)	(5, 6)	IMAGE
ϕ	-0.16537	0.28698	-0.18208	
t		1.4685	1.4868	
y	1.5	1.1357	1.2515	
u	0	-0.2481	0.0779	-0.1500
$\bar{y}$	-0.8	-0.07119	0.63633	
$\bar{u}$	0.364	0.49630	0.47587	0.36000
ν_{F-C}	60.3	36.2	60.3	
$a = -\frac{y^2\phi}{\nu}$	-0.006171	0.010226	-0.004730	$\Sigma a = -0.000675$
$b = -\bar{y}y\phi/\nu$	0.003291	-0.000641	-0.002405	$\Sigma b = 0.000245$

$$\text{TAch} = \frac{-\Sigma a}{(n_{k-1}u_{k-1})} = -0.00450$$

$$\text{Tch} = \frac{-\Sigma b}{(n_{k-1}u_{k-1})} = 0.00164$$

Table 6.13 - Thin lens computation of axial and lateral color for a triplet. ($f = 10$)

thin lenses is given by,

$$\mathrm{TAch}_{v-r} = \frac{1}{(n_{k-1} \; u_{k-1})} \sum_{j=1}^{j=\eta} \left(y^2 \frac{\phi}{\nu_{v-r}} \right)_j = \frac{-\Sigma a}{(n_{k-1} \; u_{k-1})} , \qquad (40)$$

and

$$\mathrm{Tch}_{v-r} = \frac{1}{(n_{k-1} \; u_{k-1})} \sum_{j=1}^{j=\eta} \left(y \bar{y} \frac{\phi}{\nu_{v-r}} \right)_j = \frac{-\Sigma b}{(n_{k-1} \; u_{k-1})} . \qquad (41)$$

The use of these equations is illustrated in Table 6.13. The system used in the table is very close to the thin lens equivalent of the system shown in Table 6.7. Note how the angle u of the axial ray as it passes through the system is the same, to four decimal places, for both examples. The TAch for the equvalent lens is not exactly the same as for the thick lens due to the thicknesses of the elements.

6.10.8 Thin lens achromatic system.

6.10.8.1 If Equation (40) is written for two closely spaced lenses (a) and (b), and combined, there results

$$\mathrm{TAch}_{v-r} = \frac{1}{(n_{k-1} \; u_{k-1})} \left[y^2 \left(\frac{\phi}{\nu_{v-r}} \right)_a + y^2 \left(\frac{\phi}{\nu_{v-r}} \right)_b \right] . \qquad (42)$$

This is an expression for the axial chromatic aberration of the doublet lens. In order to make $\mathrm{TAch}_{v-r} = 0$, it is necessary that,

$$\left(\frac{\phi}{\nu} \right)_a = \left(- \frac{\phi}{\nu} \right)_b . \qquad (43)$$

In Table 6.9 it was shown that for two thin lenses in contact,

$$\phi = \phi_a + \phi_b .$$

Combining this equation with Equation (43) yields the relations,

$$\phi_a = \phi \frac{\nu_a}{\nu_a - \nu_b} , \qquad (44)$$

and

$$\phi_b = - \phi \frac{\nu_b}{\nu_a - \nu_b} . \qquad (45)$$

6.10.8.2 Equations (44) and (45) enable one to pick two glasses with different ν - values and calculate the powers of the two lenses to make an achromatic lens. It is important to realize that these equations reduce the transverse axial chromatic aberration to zero only for the two wavelengths λ_v and λ_r. These are the two wavelengths used to compute the value of ν for the glasses, where the ν - number of a glass is defined as,

$$\nu_{(v-r)} = \frac{n_g - 1}{n_v - n_r} . \qquad (46)$$

On the other hand, other wavelengths do not come to the same focus as λ_v and λ_r. The chromatic aberration TAch_{v-g} between an intermediate wavelength λ_g and λ_v may be calculated by substituting $\nu_{v-g} = \frac{n_g - 1}{n_v - n_g}$ for each element and inserting them in Equation (42). Then

$$\mathrm{TAch}_{v-g} = \frac{1}{(n_{k-1} \; u_{k-1})} \left[y^2 \left(\frac{\phi}{\nu_{v-g}} \right)_a + y^2 \left(\frac{\phi}{\nu_{v-g}} \right)_b \right] . \qquad (47)$$

Since the lens was adjusted to be an achromat for λ_v and λ_r then Equation (43) must also be satisfied. This equation can be readily inserted in Equation (47) by an obvious redefining of ν_{v-g}, as follows,

$$\nu_{v-g} = \left(\frac{n_g - 1}{n_v - n_g} \right) \left(\frac{n_v - n_r}{n_v - n_r} \right) = \nu_{v-r} \left(\frac{n_v - n_r}{n_v - n_g} \right) .$$

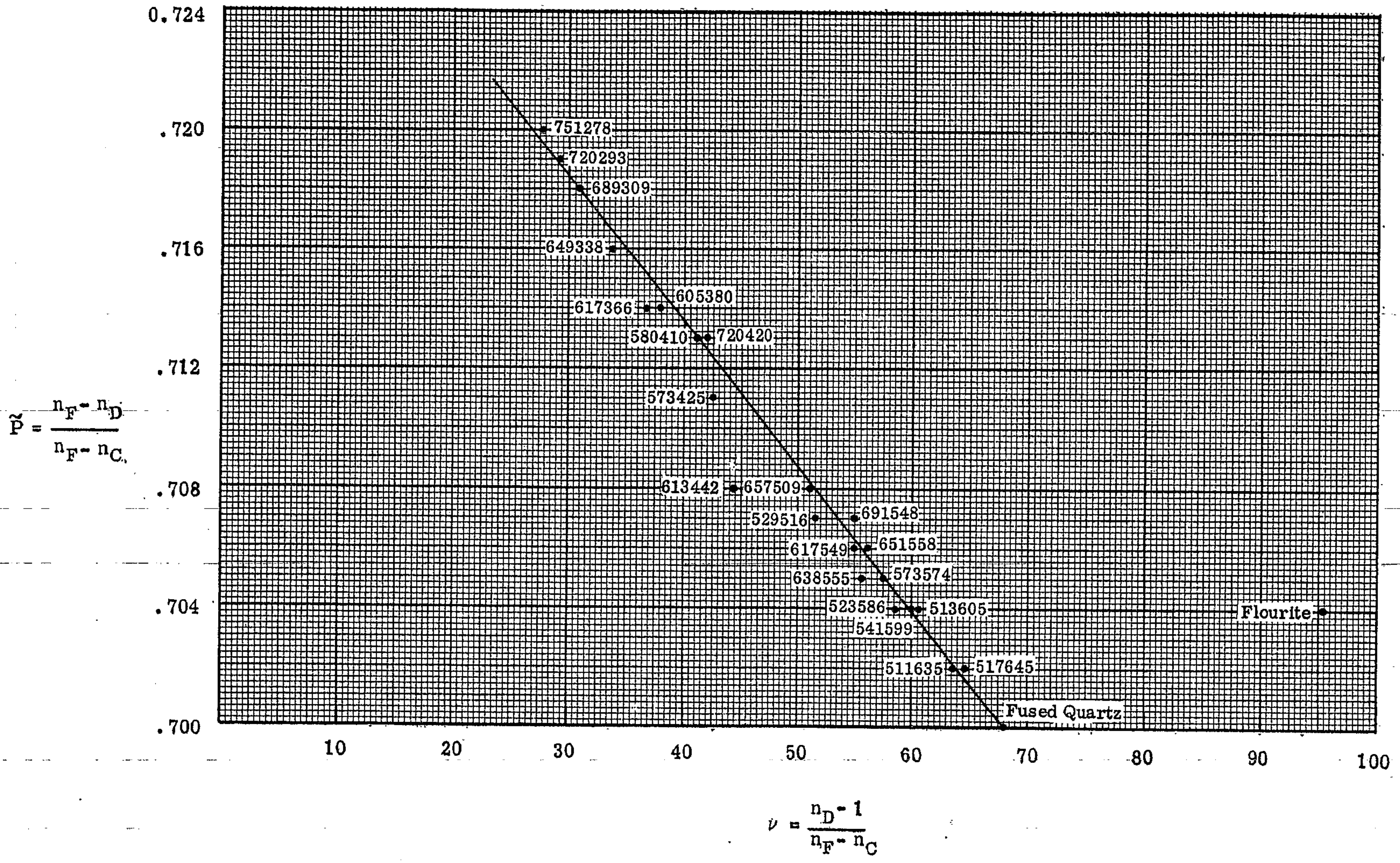


Figure 6.18 - $\tilde{P}$ versus ν_{F-C} for several glasses

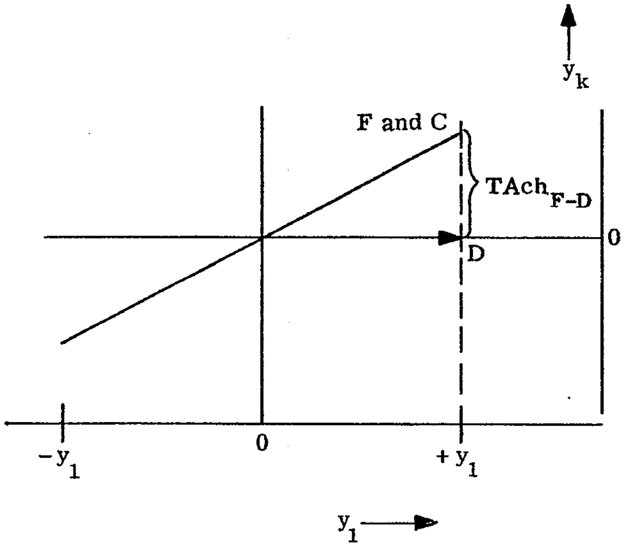


Figure 6.19 - Transverse axial chromatic aberration for an achromatic objective corrected for F and C light.

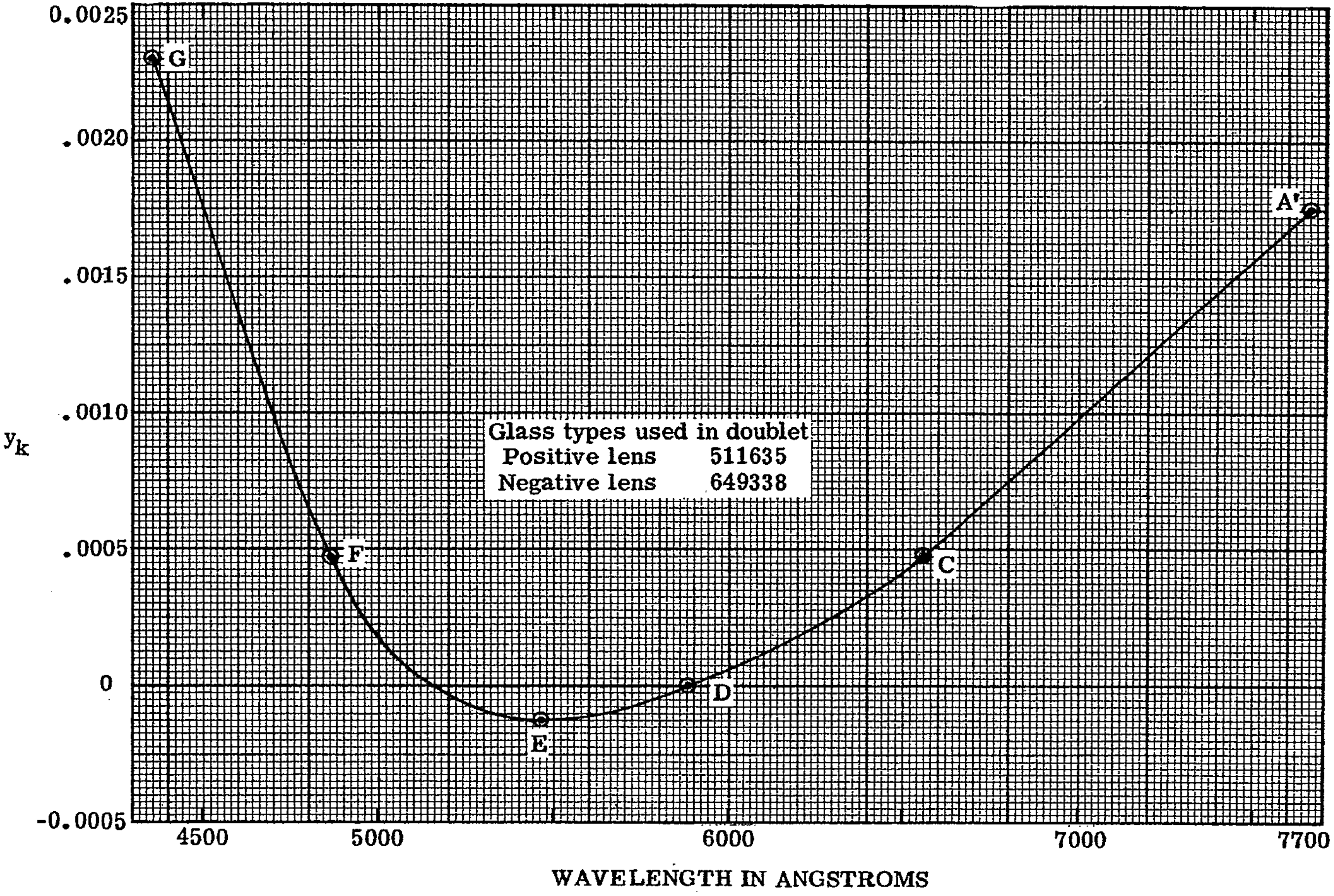


Figure 6.20 - Plot of y_k versus λ for an achromatic doublet.

Defining the partial dispersion ratio (see Paragraph 2.7.3),

$$\tilde{P} = \frac{(n_v - n_g)}{(n_v - n_r)} ,$$

we have

$$\nu_{v-g} = \nu_{v-r} / \tilde{P} . \qquad (48)$$

Equation (47) then becomes, with the help of Equations (44) and (45),

$$\mathrm{TAch}_{v-g} = \frac{-y}{n_{k-1}} \left[\frac{\tilde{P}_a - \tilde{P}_b}{\nu_a - \nu_b} \right] . \qquad (49)$$

6.10.8.3 Equation (49) gives the value of the transverse aberration between λ_v and λ_g, when λ_v and λ_r wavelengths are united. The equation indicates that if λ_v, λ_g, and λ_r are to be brought to focus simultaneously, then $\tilde{P}_a = \tilde{P}_b$. Most glass catalogs give values of $\tilde{P}$ for many combinations of wavelength for each glass. In Figure 6.18 the value of $\tilde{P}$ for $\frac{F - D}{F - C}$ is plotted against ν_{F-C} for several types of glass. As will be pointed out in later sections, a doublet should be designed with low powers of the individual elements. Equations (44) and (45) show that the powers of the (a) and (b) elements of a doublet may be kept small by selecting optical glasses with large differences in ν. Usually doublets should have ν differences larger than 20. As can be seen from the slope of Figure 6.18, for almost any combination of glasses one can select, the ratio of $(\tilde{P}_a - \tilde{P}_b)/(\nu_a - \nu_b)$ is a constant equal to $-1/2200$. When this number is substituted into Equation (49), TAch_{v-g} is positive for positive y. Reference to Figure 6.13 indicates that for positive TAch_{v-g} the axial ray in D light crosses the axis closer to the lens than the axial ray in F light. Using Equation (13) and noting that $(u_{k-1})_F = (u_{k-1})_D$ to this approximation, we see that if F and C wavelengths are united, then D light focuses closer to the lens by the amount $f'/2200$, if the lens is in air. In Figure 6.19 a plot similar to that of Figure 6.14 is shown for a typical achromatic doublet, corrected to unite F and C light. It is instructive to plot the transverse axial aberration as a function of wavelength. This has been done in Figure 6.20 for an achromatic lens. Note how the curve has a minimum near $\lambda = 5500A$. This is the wavelength at the peak of sensitivity for the eye, which is the reason F - C achromatism is considered to be proper for visual systems.

6.10.8.4 TAch_{F-D} is called the secondary spectrum or the secondary color. It is a very difficult aberration to eliminate with ordinary glass types, and often sets the limiting aperture for a lens. The following methods may be used to reduce the secondary spectrum in a lens system.

(1) Use special materials with equal partial dispersions.

(2) Use more than two types of glass.

(3) Use proper combinations of lenses.

More information on the correction of the secondary spectrum will be given in Section 11 under the design of telescope objectives. One can use Equation (40) to compute the secondary color for more complex optical systems, such as air spaced doublets, triplets, or combinations of doublets: however, the algebra becomes so complicated that it is difficult to obtain useful equations like (49) for anything more complicated than a closely packed doublet. It can be shown however, that for a given pair of glasses, the secondary color increases as the air space increases. The relation between secondary spectrum and separation of the two elements is derived by a method similar to that used for Equation (49). First an equation analogous to Equation (42) is derived; this will involve the separation of the elements as well as the powers and ν - numbers. The condition for C - F achromatism, analogous to Equation (43) is then found. The total power for a dialyte from Table 6.9 is used with the achromatic condition to find the analogs of Equations (44) and (45). By the method given in 6.10.8.1, the equations analogous to (47) and (49) are then derived.

6.10.8.5 Although Section 6 deals with first order optics, and hence with the chromatic aberrations, we will mention here one of the third order aberrations, Petzval curvature, because of its close connection with the secondary spectrum. Petzval curvature, known also as curvature of field, has the following physical meaning. For monochromatic light, if spherical aberration, coma, and astigmatism are absent, the point images of point objects lie on a surface, generally curved. Near the optical axis this surface can be considered spherical with a curvature called the Petzval curvature. Flat-field systems have zero or very small Petzval curvature.

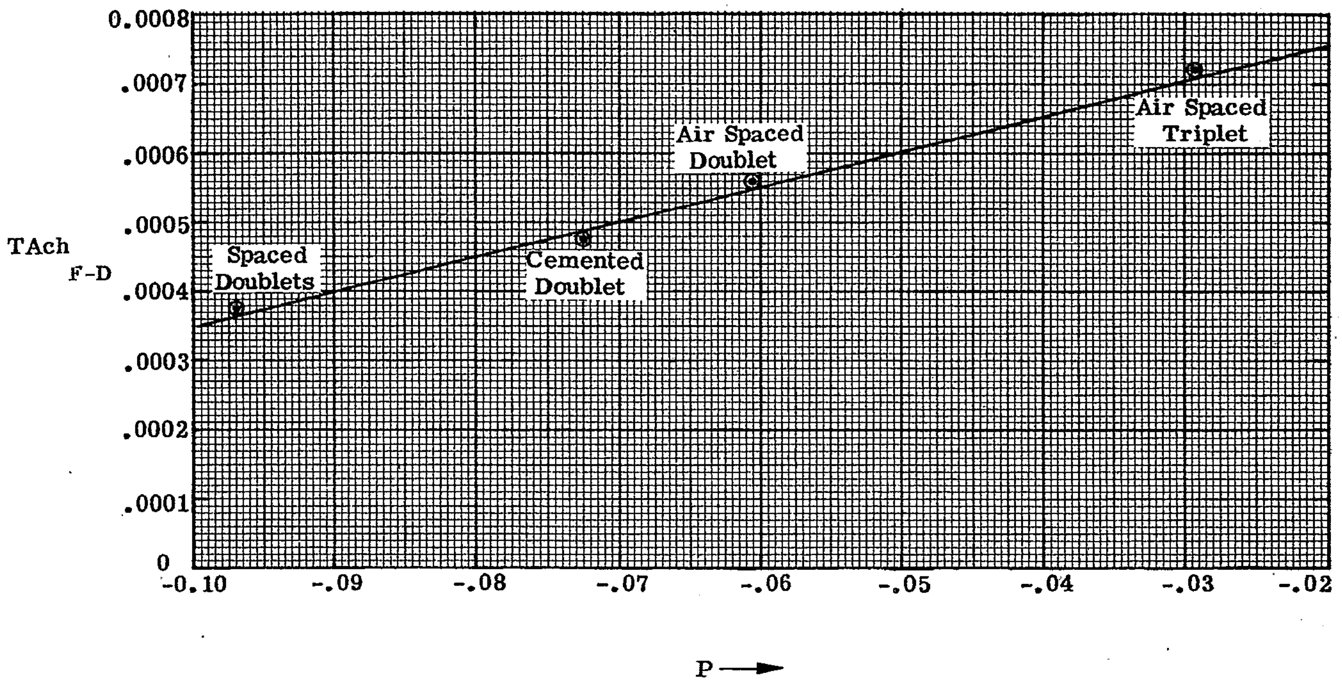


Figure 6.21 - Plot shows qualitative connection between Petzval curveature and secondary color. (Image is assumed in air).

6.10.8.6 Section 8 will discuss how the Petzval contribution for each surface is calculated. When the two surface contributions for a simple lens are added, as was done for the chromatic contributions in Paragraph 6.10.7.1, the Petzval contribution of a simple lens is $P = -\phi/n$. For a system of thin lenses in air, $-P$ is the sum of the power, ϕ, divided by the index for each lens. $P = -\sum_{j=1}^{j=n} \left(\frac{\phi_j}{n_j}\right)$. If the $TAch_{F-D}$ is plotted versus P for lens types, the points lie along an approximate straight line. This is shown in Figure 6.21. To obtain the data for this curve, a zero spaced doublet, an air spaced doublet, a positive-negative-positive-triplet, and two widely spaced achromatic doublets (a Petzval lens) were set up for computation. Each system has an exact focal length of 10 and is corrected for zero $TAch_{F-C}$. The axial paraxial ray was traced through at $y = 1.0$. All the positive lenses were of 511635 glass and all the negative lenses were of 649338 glass. This approximately linear relationship causes real difficulty in the design of flat-field lenses, since reduced Petzval curvature tends to accompany an increase in the amount of secondary color. This is a particularly serious problem in the design of periscope systems.

6.11 ENTRANCE AND EXIT PUPILS, THE CHIEF RAY AND VIGNETTING

6.11.1 General. As shown in Section 6.4, the complete analysis of the first order properties of an optical system can be found by tracing two rays through the optical system. Any two rays may be used, but it is convenient to pick the two rays with some care. In order to specify quantitatively which two rays are usually used, we must discuss the meanings of the pupils of an optical system.

6.11.2 The aperture stop. The bundle of rays, which proceed from an object point to the image point through an optical system, is limited in the sense that all the rays in the entire solid angle of 4π sterradians do not get through the system. The aperture stop is the physical stop or diaphragm, as distinguished from an image of a stop, which limits the rays passing through the system. The aperture stop may be a lens or it may be an opening in an otherwise opaque surface. It is almost always circular; we will consider it as such since we are concerned with systems having rotational symmetry.

6.11.3 Entrance and exit pupils.

6.11.3.1 The pupils are images of the aperture stop. The entrance pupil is the image of the aperture stop in the part of the system preceding the aperture stop. Hence to locate the entrance pupil, given the position of the aperture stop, an axial paraxial ray is traced backwards through the system from the center of the aperture stop. The point where it last crosses the axis is the entrance pupil point. The entrance pupil plane is a plane perpendicular to the axis at the entrance pupil point. If the diameter of the aperture stop is known, an oblique paraxial ray is traced backwards from the rim or margin of the aperture stop. The intersection of this ray with the entrance pupil plane gives the radius of the entrance pupil.

6.11.3.2 Similarly, the exit pupil is the image of the aperture stop in that part of the system following the aperture stop. By tracing an axial and oblique ray from the aperture stop, the exit pupil plane can be located, and the diameter of the exit pupil can be determined. It sometimes happens that the aperture stop precedes (or follows) the rest of the system. In this case the aperture stop coincides with the entrance pupil (or exit pupil).

6.11.4 The chief ray. The chief ray is an oblique ray from an off-axis object point, which intersects the axis at the entrance pupil point, the center of the aperture stop, and the exit pupil point. Because it passes through the centers of the pupils and the aperture stop, it is approximately the central ray of the conical bundle from the object point to the image point. Hence it is representative of the entire bundle.

6.11.5 Two convenient paraxial rays. The usual procedure is to trace one ray from the point on the object plane intersected by the optical axis ($y_o = 0$). The angle with the optical axis, u_o, should be chosen to equal one half the actual cone angle to be passed by the optical system. Hence this ray passes through the margin of the pupils and the aperture stop. u_o is the radius of the entrance pupil divided by the distance between object surface and entrance pupil plane. The second ray should be traced from a point $\bar{y}_o$ in the object plane corresponding to an object near the maximum size to be accommodated by the lens system. This second ray is a chief ray from the object point chosen. Hence $\bar{u}_o$ is $\bar{y}_o$ divided by the distance between object surface and entrance pupil plane. (See Figure 6.22).

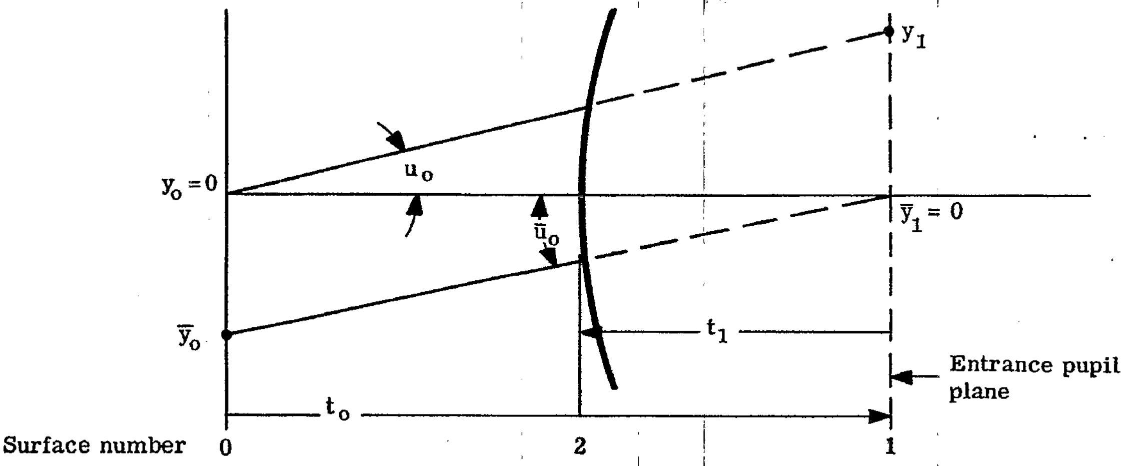


Figure 6.22 -Location of entrance pupil and numbering of surfaces.

6.11.6 Pupils as surfaces in the optical system. Many designers include the entrance pupil plane as a plane surface in the system. It is labelled number one. The actual first surface of the lens may be encountered before the chief ray reaches the entrance pupil. In this case the thickness t_1 is made negative, indicating the entrance pupil plane is actually virtual. As the chief ray passes through the lens it may cross the axis at several positions. Each position is called an aperture plane. After it finally emerges in the image space it can be extended until it crosses the axis. This position is the exit pupil plane of the system and is numbered the (k - 1) surface. Although it is not necessary to include the entrance and exit pupil planes in the calculations of a lens, their inclusion is helpful because they are excellent planes of reference. It is convenient to describe aberration data by using the image coordinates plotted against their conjugate coordinates in the entrance pupil. (See Section 8).

6.11.7 Numerical example. As an example of the foregoing material, Figure 6.23 shows the pupils for a two-lens system. Table 6.14 shows the calculations for this system. In the example, the entrance pupil plane is found in the following way. As the chief ray is drawn, the lens (a) bends it up and the lens (b) bends it back down. It is nearly always true that ($\Delta u_a + \Delta u_b$) should be close to zero. This tends to keep the distortion corrected. (Distortion is a monochromatic aberration which will be discussed in Section 8). Since $\Delta u = u - u_{-1}$, Equation (24) shows that to meet this condition, $y_a \phi_a + y_b \phi_b = 0$.

The chief ray, therefore, must cross the axis between the two lenses and divide the space in the ratio of ϕ_b / ϕ_a . A value of -1 for $\bar{y}_a$ and $+1$ for $\bar{y}_b$ may be selected for convenience in this problem, because ϕ_a and ϕ_b are equal. Since $t_2 = 5.0$, $\bar{u}_2 = 0.4$, and the chief ray can then be traced backwards to the object plane as shown in the example. The entrance and exit pupil planes are located by solving for t_1 and t_3 to make $\bar{y}_1 = \bar{y}_4 = 0$. Since $\bar{y}_a = -1$ was used for convenience, the object height may come out to be far different from the value to be used for the true object. If the designer wishes to have a ray traced from the true object height, it may be done by simply scaling all the ray data for the chief ray. In the sample $\bar{y}_o$ came out -4. A second ray was traced at $\bar{y}_o = -2$.

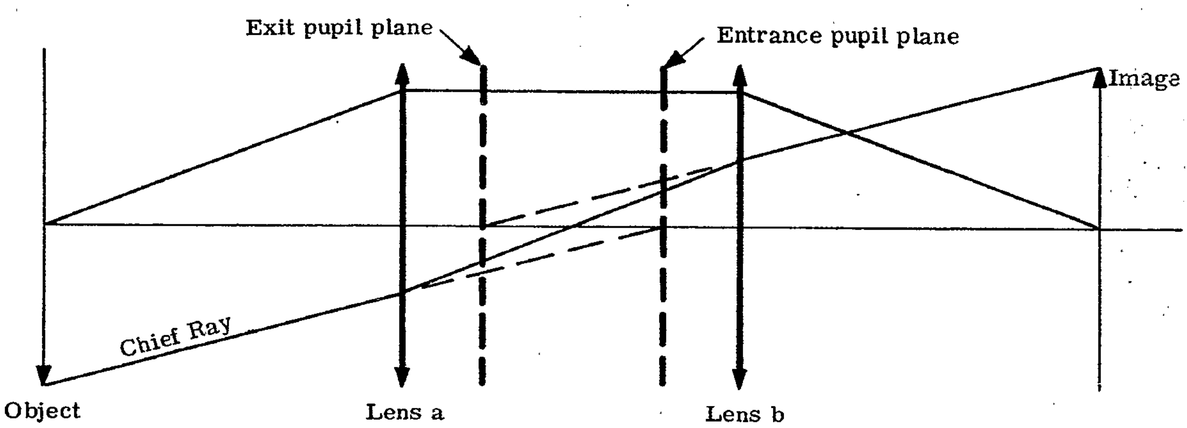


Figure 6.23 - Illustration of entrance and exit pupils.

Surface	Object Plane 0		Entrance Pupil Plane 1		Lens (a) 2		Lens (b) 3		Exit Pupil Plane 4		Image Plane 5
$-\phi$	0		0		-0.1		-0.1		0		0
t		13.33		-3.33		5		-3.33		13.33	
y	0		1.33		1		1		1.33		0
u		0.1		0.1		0		-0.1		-0.1	
$\bar{y}$	-4		0		-1		1		0		4
$\bar{u}$		0.3		0.3		0.4		0.3		0.3	
$\bar{y}$	-2		0		-0.5		0.5		0		2
$\bar{u}$		0.15		0.15		0.2		0.15		0.15	

Table 6.14 - Calculations showing location of entrance and exit pupil planes.

6.11.8 Vignetting.

6.11.8.1 In the above discussion on the aperture stop and pupils it was assumed that the aperture stop was circular. Hence the pupils are circular and a circular cone of rays passes through the system from an axial object point. For an off-axis object point, the cone of rays limited by the aperture stop will not be circular; and the entrance pupil will generally subtend at the object point a smaller solid angle than for an axial object point. This phenomenon is called vignetting; the oblique bundle of rays is said to be vignetted.

6.11.8.2 In the example shown in Table 6.14 the path of the chief ray has been calculated through a simple two-element lens. The next question is, what is the shape of the beam of light that passes through the optical system from the oblique object point ? To answer this, it is necessary to project all the lens apertures in the system onto the entrance pupil plane. Since the path of any ray can be readily computed as a linear combination of two rays (see Section 6.4), it is possible to compute the coordinates in the entrance pupil plane for any ray from the object point of interest which passes through any part of any aperture of the system. For example, suppose we wish to find the coordinate on the entrance pupil plane of a ray from the object $\bar{y}_o = -2$, which passes through the center of the (a) lens. Since two rays have been traced through the lens, a value of y and $\bar{y}$ is known on each surface. Any other ray $\bar{\bar{y}}$ may be traced from the object point y_o with the use of Equation (10a)

$$A\,\bar{y}_j + B\,y_j = \bar{\bar{y}}_j \;.$$

On the object plane $y_o = 0$, $\bar{y}_o = \bar{\bar{y}}_o$. Therefore

$$A = \bar{\bar{y}}_o / \bar{y}_o = 1 \,,$$

and for the i th surface,

$$B = \frac{\bar{\bar{y}}_i - \bar{y}_i}{y_i} \;.$$

Finally then,

$$\left[\, \bar{\bar{y}}_j - \bar{y}_j \,\right] = \left[\, \bar{\bar{y}}_i - \bar{y}_i \,\right] \frac{y_j}{y_i} \;. \tag{50}$$

To calculate the coordinate of any ray on the entrance pupil plane, which has the coordinate $\bar{\bar{y}}_i$ on the i th surface, Equation (50) becomes

$$\bar{\bar{y}}_1 = (\,\bar{\bar{y}}_i - \bar{y}_i\,)\,\frac{y_1}{y_i} \,,$$

since $\bar{y}_1 = 0$.

6.11.8.3 In the example shown in Figure 6.23 and Table 6.14, a ray from the object point $\bar{y}_o = -2$ passing through the center of the (a) lens ($\bar{\bar{y}}_2 = 0$) will project onto the entrance pupil plane at the value $\bar{\bar{y}}_1 = (0 + 0.5)(1.33)/1 = 0.666$. The top edge of the (a) lens (assumed $\bar{\bar{y}}_2 = 1$) will appear in the entrance pupil plane at $\bar{\bar{y}}_1 = (1 + 0.5)(1.33) = 2$. The center of the (b) lens will project in the entrance pupil plane at $\bar{\bar{y}}_1 = (0 - 0.5)(1.33) = -0.666$. The top edge of the (b) lens (assume $\bar{\bar{y}}_3 = 1$) will appear in the entrance pupil plane at $\bar{\bar{y}}_1 = (1 - 0.5)(1.33) = 0.666$.

6.11.8.4 Since the center and top edge of each lens, (a) and (b), are now projected on the entrance pupil plane, it is possible to construct circles indicating the complete aperture of the lenses as they appear in the entrance pupil plane. These apertures are shown in Figure 6.24. Only those rays passing through the area common to both circles will pass through the two lenses. In order to have the same aperture for the oblique beam as for the central beam, an aperture would have to be placed to appear as the inner circle shown in Figure 6.24. A circular aperture in the entrance pupil plane of radius 0.666 just fits in the common area of the two circles. Now in this case, the entrance pupil plane is virtual, so no physical stop can be placed in it. Since the chief ray does actually cross the axis at a point midway between the lenses, the physical aperture stop may be placed in this position and it will appear as a central stop in the entrance pupil plane. Using Equation (50) , the size of the aperture stop can be calculated using the following data.

$\bar{\bar{y}}_1 = 0.666 =$ height of edge of entrance pupil aperture.

$\bar{\bar{y}}_i =$ height of edge of aperture stop in the aperture plane.

$\bar{y}_i = 0 =$ height of chief ray in the aperture stop plane.

$y_i = 1.0 =$ height of axial ray in the aperture stop plane.

$y_1 = 1.33 =$ height of axial ray in the entrance pupil plane.

Therefore

$$\bar{\bar{y}}_i = \frac{0.666}{1.33} = 0.5 \,.$$

6.11.8.5 Usually some vignetting for the oblique beams is allowed, so the aperture stop is made larger than the largest circle included in the common area. Figure 6.25 shows the appearance of the aperture stop when it is made 0.75 in radius. The clear area is the common area for all the apertures, and its area is a measure of the total light passing through the system from the oblique object point. The common area is 67 % of the area of the image of the (0.75) aperture stop in the entrance pupil plane. Therefore, the oblique beam is vignetted by 33 % . All other factors remaining constant, the illumination at the image point, $\bar{y}_k = 2$, is 67 % of the illumination at the point $y_k = 0$. In Figure 6.25, the aperture stop of radius 0.75 located midway between the (a) and (b) lens, is imaged in the entrance pupil plane with a radius of 1.0 . The exit pupil also has a radius of 1.0 .

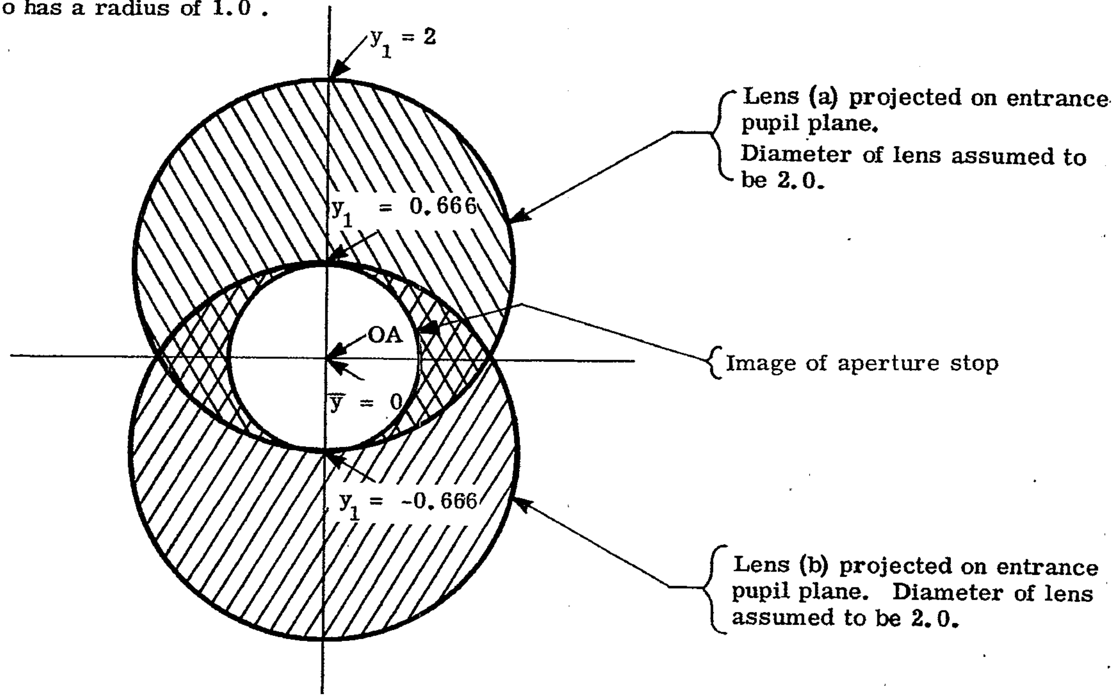

$y_1 = 2$
Lens (a) projected on entrance pupil plane.
Diameter of lens assumed to be 2.0.
$y_1 = 0.666$
OA
$\bar{y} = 0$
Image of aperture stop
$y_1 = -0.666$
Lens (b) projected on entrance pupil plane. Diameter of lens assumed to be 2.0.

Figure 6.24 - Apertures of the (a) and (b) lenses and of the aperture stop projected onto the entrance pupil. The oblique beam is not vignetted.

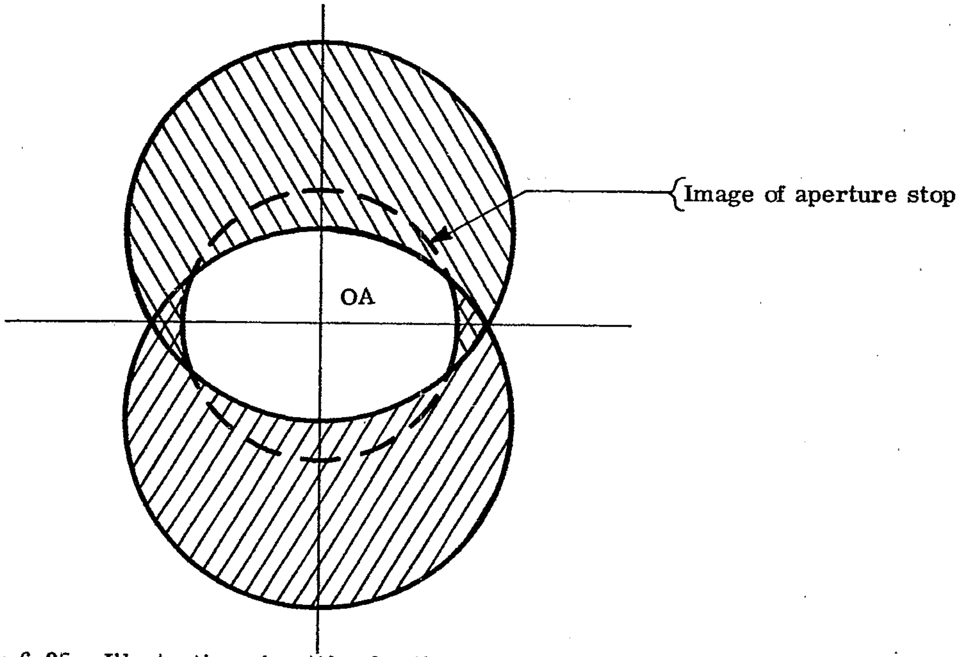


Figure 6.25 - Illustrating vignetting for the same system shown in Figure 6.24 but with a larger aperture stop.

7 SIMPLE THIN LENS OPTICAL SYSTEMS

7.1 INTRODUCTION

7.1.1 Thin lens solution. The first two steps in designing optical systems, which will be discussed further in Section 9, are (1) selecting lens types for the various elements, and (2) finding a first order solution assuming thin lenses. The methods and procedures used in step (2) were developed in Section 6. In Section 7 the optics of several thin lens optical systems will be described to illustrate the usefulness of the paraxial equations and to indicate the reasoning a designer uses in following step (1). With information obtainable from only paraxial ray data, a designer can conclude many of the important features needed for a final design. There are numerous discussions in text books of simple optical systems such as the microscope and the telescope. The following discussion, assuming that the reader has read some explanation of these systems, will concentrate on the numerical analysis.

7.1.2 Optical systems used with the eye. The optical systems considered in this section are all used with the human eye. Because of this the eye is an integral part of the system and must be considered in the design. There are four basic types of lenses: (1) microscope objectives, (2) telescope objectives, (3) eyepieces and (4) photographic objectives. The first three are used with the human eye and systems employing these types are discussed in Section 7.

7.2 THE SIMPLE MAGNIFIER

7.2.1 A single lens. One of the simplest of optical devices is the simple magnifier. A single positive lens works as a magnifier because it makes an object appear to subtend a larger angle at an observer's eye than is possible with the unaided eye. Without a magnifier, an observer can make an object appear larger only by bringing it close to his eye. As an object is moved closer and closer to an observer's eye it is necessary for the eye to increase its refractive power in order to continue to focus the image on the retina. There is a minimum distance V at which the eye has increased its refractive power to its maximum capability. For object to eye distances less than V the image will no longer be sharply focussed on the retina. For the standard observer this distance V is approximately 10 inches or 250 mm (V is always considered positive). Therefore, in order to make the object appear still larger, it is necessary to add refractive power to the eye, so that the object may in effect be brought closer. The magnifier provides the extra refractive power required.

7.2.2 Magnifying power. The magnifying power of a visual instrument may be defined as

$$MP = \frac{\text{size of retinal image obtained with instrument}}{\text{size of retinal image obtained with the unaided eye}} .$$

In the region of the paraxial approximation this is equivalent to the definition, $MP = \beta/\alpha$ where β equals one half the angle subtended by the object as seen through the instrument, and α equals one half the angle subtended by the object as seen by the naked eye. (These particular definitions of α and β assume that the object and image are centered with respect to the optical axis. This is usually the case with visual instruments. According to this assumption, when reference is made to "an object $\bar{y}_o$," the object height is rigorously $2\bar{y}_o$.) β is called the half image field angle, and α is called the half object field angle. Magnifying power, then, is the ratio of the field angles.

7.2.3 Diagram of a single lens magnifier. In Figure 7.1 (a) an object $\bar{y}_o$ is shown viewed with the unaided eye. Figure 7.1 (b) shows the same object viewed by a single lens magnifier. The eye is placed at a distance d from the lens. The object is placed in relation to the lens so that the visual image $\bar{y}_k$ lies to the left of the eye at a distance, A. (A is negative). For the eye to focus on the image, A must be numerically equal to or larger than V. The numerical formulation of this problem may be handled with simplicity by the usual methods adequately covered in most text books. In the following analysis, it will be handled formally using the ray trace format in order to illustrate a method of analysis which can be used for any system, regardless of its complication.

7.2.4 Ray trace format.

7.2.4.1 The system consists of an object plane, an entrance pupil plane, a thin lens, an aperture stop and exit pupil, and a final image plane. Table 7.1 contains a layout of a computation sheet for this simple magnifier system. The data may be filled in as follows. First, all the ϕ values are zero except that of the lens. We also know that $y_o = 0$ and we may choose to trace a paraxial ray at any angle from the point $y_o = 0$.

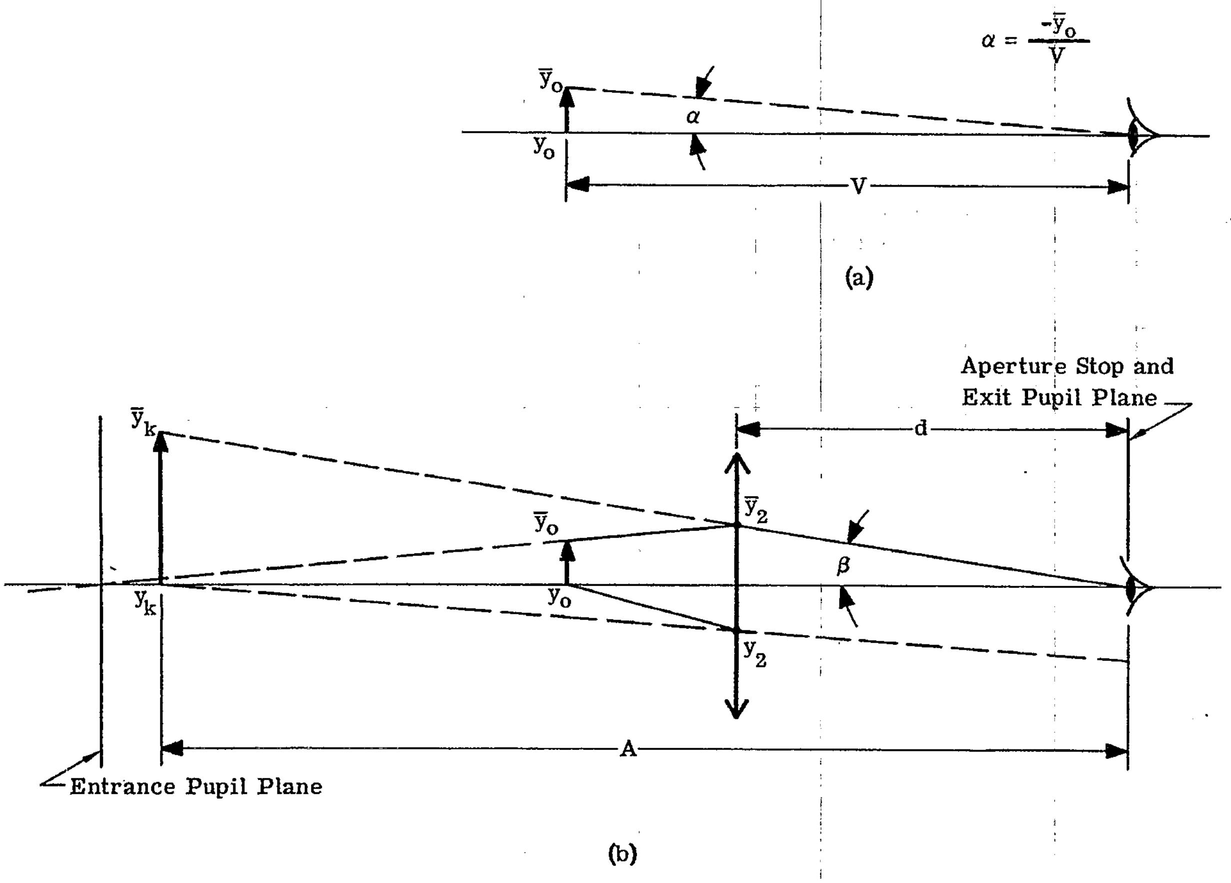


Figure 7.1 - Diagrammatic illustration of a single lens magnifier.

Surface	Object 0	Entrance Pupil Plane 1	Lens 2	Exit Pupil Plane 3	Image Plane 4
$-\phi$	0	0	$-\phi$	0	0
t	$-\bar{y}_0/\beta(1-d\phi)$	$-d/(1-d\phi)$	d	A	
y	0	$A/(1-d\phi)(d+A)$	1	$A/(d+A)$	0
u	$\left[\phi - \frac{1}{d+A}\right]$	$\left[\phi - \frac{1}{d+A}\right]$	$-1/(d+A)$	$-1/(d+A)$	
$\bar{y}$	$\bar{y}_o$	0	$-d\beta$	0	$\bar{y}_k$
$\bar{u}$	$\beta(1-d\phi)$	$\beta(1-d\phi)$	β	β	

Table 7.1 - Computation sheet for a simple magnifier

Therefore, we elect to make $y_2 = 1$. (Figure 7.1 (b) shows y_2 negative; this has been so done for pictorial clarity). The system has only two physical stops, the lens and the eye pupil. One of these is the aperture stop. To find which one, we can image each stop in the system preceding it, and see which image subtends the smallest angle at the base of the object, $y_o = 0$. The image of the lens in the lens is of magnification unity and is at the lens; hence if the size of the lens and its distance from the object is known, the angle subtended can be found. Likewise if the eye pupil size and location is known, its image in the lens can be determined, and the angle subtended by this image compared with the previous angle. We see, therefore, that which of the two stops is the aperture stop depends on the size and location of the two elements, i.e. on the design of the system. In such systems it is usual to assume that the lens is so much larger than the eye pupil that the latter is the aperture stop. Hence it is also the exit pupil. Therefore we can fill in d, the distance from the lens to the aperture stop plane, and A the distance from the aperture stop plane to the final virtual image plane. Since $y_k = 0$ and $u_2 = u_3$ (no refraction occurs at surface 3), it can be concluded that $u_2 = u_3 = -1/(d + A)$. With y_2 and u_2 known it is possible to calculate u_1 from equation 6-(24). Then $u_1 = \phi - 1/(d + A)$. Also $u_o = u_1$.

7.2.4.2 The oblique principal ray may now be traced backward through the system from the center of the exit pupil. Let this go back at the angle β with the optical axis. $\bar{y}_2$ is now determined as $-\beta d$. $\bar{u}_1$ and $\bar{u}_o$ are also determined. Knowing $\bar{y}_o$, $\bar{y}_1$, $\bar{y}_2$, $\bar{u}_o$ and $\bar{u}_1$, t_o and t_1 may be computed. Since all the spaces are now determined y and u are known on every surface for each ray.

From Equation 6-(7),

$$\bar{y}_k = \bar{y}_o \frac{u_o}{u_{k-1}} .$$

Therefore,

$$\bar{y}_k = \bar{y}_o \left[1 - \phi (d + A) \right] .$$

Since

$$\beta = \frac{\bar{y}_k}{A} \quad \text{and} \quad \alpha = - \frac{\bar{y}_o}{V} ,$$

$$MP = - \frac{V}{A} \left[1 - \phi (d + A) \right] . \tag{1}$$

7.2.5 Analysis of magnifying power equation. There are several cases of special interest which should be noted.

a) If $A = -\infty$

$MP = V\phi$

b) If $d = f' = \frac{1}{\phi}$

$MP = V\phi$

c) If $A = -V$ with $d = 0$

$MP = 1 + V\phi$.

One can see by inspection of these equations, that $MP = V/f'$ is the minimum magnifying power and $MP = (V/f') + 1$ is the maximum. Hence for maximum magnifying power, the final image is at the near point of the eye, a distance V from the eye. Therefore the eye has maximum refractive power. For the relaxed eye, the image is at the far point; this is ∞ for the normal eye and results in a minimum magnifying power. The relative increase in magnifying power, as the eye accommodates for smaller A, is small and offset by the greater likelihood of eye strain. (For a typical magnifier of 1 inch focal length, the maximum magnifying power is only 10% higher than the minimum). Therefore simple magnifiers should be designed and used so that the final image is at infinity, or at the far point of the eye, if these cases differ.

7.3 THE MICROSCOPE

7.3.1 Limitations of a simple magnifier. It is clear from Section 7.2.5, that for large magnifying power, ϕ must be large, hence the focal length, f', must be small. Because the final image is to be at infinity, the object must be at F_1. Therefore for large magnifying power, the object must be placed very close to the lens. By Equation 6-(22) we see that the lens surfaces must have very short radii, and therefore the diameter of the lens will be small. Because it was assumed that the eye pupil was the aperture stop, for the case of the simple magnifier, the only other stop in the system, namely the lens, is the field stop. Whereas the aperture stop limits the bundle of rays traversing the system, the field stop (a physical stop) limits the field of view. Hence a small diameter magnifying lens means a small object field.

7.3.2 The simple microscope. A practical method of overcoming the limitations of the simple magnifier is to use a relay lens as shown in Figure 7.2. While the object is being relayed it may also be magnified. The magnifying power of the microscope is then the product of the lateral magnification of the objective and the magnifying power of the eyepiece. As with the magnifier, it is advisable to adjust the microscope so that the final virtual image is at ∞; the microscope is then in afocal adjustment. In this case the eyepiece magnifying power is $MP_e = V/f'_e$, and the magnifying power of the microscope may be written

$$MP = m_o \frac{V}{f'_e} \qquad (2)$$

where m_o is the lateral magnification of the objective. The focal length of the objective can, in principle, have any value. If the focal length is made long, the overall length of the system will also be long.

7.3.3 Paraxial ray trace. Table 7.2 contains the paraxial ray trace for a microscope with an objective focal length of 16 mm and an eyepiece focal length of 25 mm. In order to use the objective lens symmetrically, i.e. in order that the chief ray pass through the center of the lens, the entrance pupil is placed in contact with the objective. The axial ray is traced from the object at an angle of 0.25. This corresponds to the sine of the angle of the actual ray to be traced through the system. This paraxial ray then passes through the optical system at very nearly the same heights as an actual true ray. As discussed in Section 23, the resolving power of a microscope depends on the wavelength and a quantity called the numerical aperture, or N.A. The numerical aperture $= n_o \sin U_o$. Since the object space has an index of $n_o = 1$, the system, as laid out, has a numerical aperture of 0.25. The chief ray was traced through the entrance pupil from an object height $\bar{y}_o = 1.0$. From this trace the exit pupil is found to lie 28.55 to the right of the lens (b).

7.3.4 Aperture stop and pupils. It is now possible to gather information about the pupils of this system from these paraxial rays. One can read directly from the table that the radius of the exit pupil is 0.625. Since the calculations are made in millimeters, the exit pupil is therefore 1.25 mm. In order that this exit pupil be the true exit pupil of the system, it is necessary to have the pupil of the observer's eye located fairly centrally in this exit pupil plane. Since the normal eye pupil is approximately 2 mm in diameter, the microscope exit pupil will definitely be the exit pupil of the entire microscope - eye system. The objective is the aperture stop and the entrance pupil of the system.

7.3.5 The f - number. The f - number of a lens (always considered positive) is defined as f/D where D is the diameter of the lens. It is very useful to calculate this quantity because from it one can estimate the difficulty of optically correcting the lens for image errors. Equation 6-(24) gives a relation between f and y for a thin lens. From this equation then, assuming $y = D/2$,

$$f\text{ - number} = 0.5/|\Delta u| .$$

In Table 7.3 the f - numbers for the objective and eyelens are listed for both the axial and oblique rays.

7.3.6 Difficulties in designing the eyepiece. The lenses should have sufficient diameter for the smallest f - number in each case. Therefore the objective should have an f - number of 1.82 and the eyepiece an f - number of 1.09. Figure 7.3 (a) shows a picture of an $f/1$ lens. This is an extremely fat lens and the chief ray would strike the curved surfaces at very large angles of incidence. The large angles of incidence introduce appreciable aberrations and the paraxial assumptions no longer hold. Hence the image at $\bar{y}_k$ would not be near the position predicted by first order theory. This lens is also uncorrected for color. Since correcting for color has the effect of approximately doubling the power of the positive element if the total power is to remain constant - because of the necessary addition of a negative element - one can see that it would be out of the question to color correct this lens. Therefore, it is clear that it will not be practical to use a single lens eyepiece. Either the size of the object will have to be reduced considerably, thereby reducing u_4 and hence increasing the f - number, or several lenses will have to be used for the eyepiece.

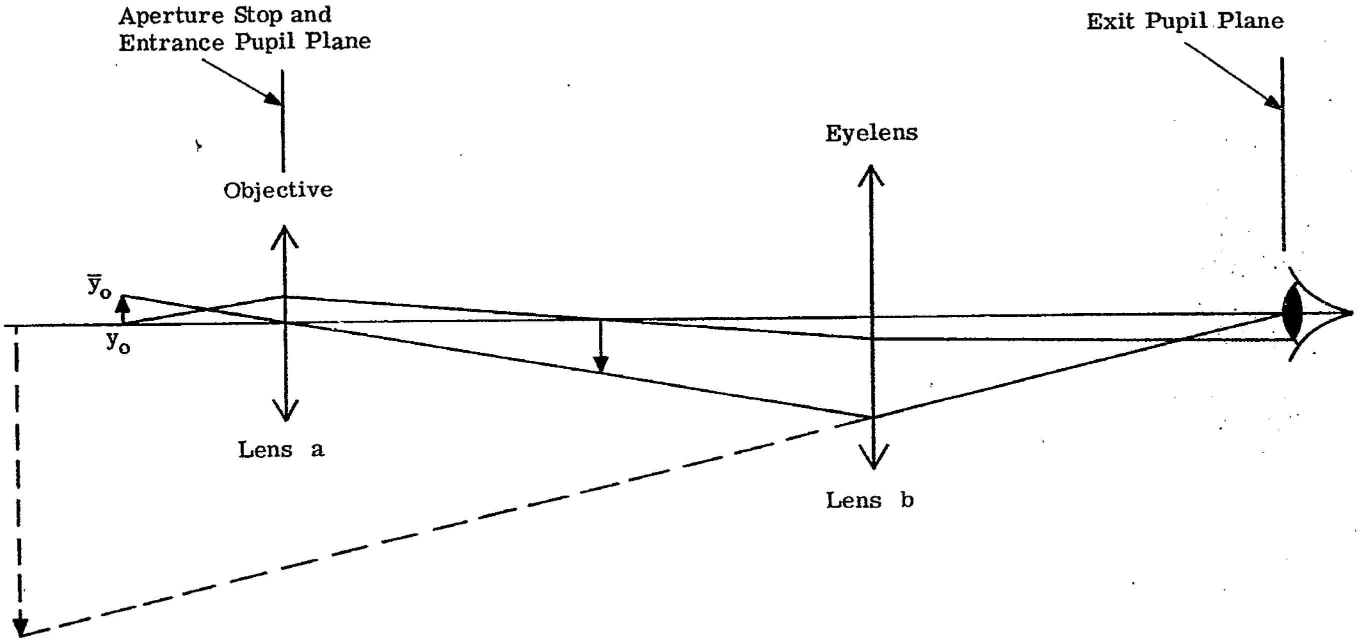


Figure 7.2 - The simple microscope. Lens (a) is the relay lens.

Surface	Object Plane 0	Entrance Pupil 1	Objective (a) 2	First Image Plane 3	Eyelens (b) 4	Exit Pupil Plane 5
$-\phi$	0	0	-0.0625	0	-0.04	
t	17.6	0	176.0	25.0	28.55	
y	0	4.4	4.4	0	-0.625	-0.625
u	0.25	0.25	-0.025	-0.025	0	
$\bar{y}$	1.0	0	0	-10.00	-11.42	0
$\bar{u}$	-0.0568	-0.0568	-0.0568	-0.0568	0.400	
ν_{F-C}	∞	∞	60	∞	60	
a	0	0	-0.02017	0	-0.00026	$\Sigma a = -0.02043$
b	0	0	0	0	-0.00476	$\Sigma b = -0.00476$
α TAch						= 0.0327
α Tch						= 0.0076

Table 7.2 - Calculations on a simple afocal microscope. All lengths are in mm.

	Axial Ray	Oblique Ray
Objective Lens	1.82	∞
Eyelens	20	1.09

Table 7.3 - f - numbers of lenses shown in Table 7.2.

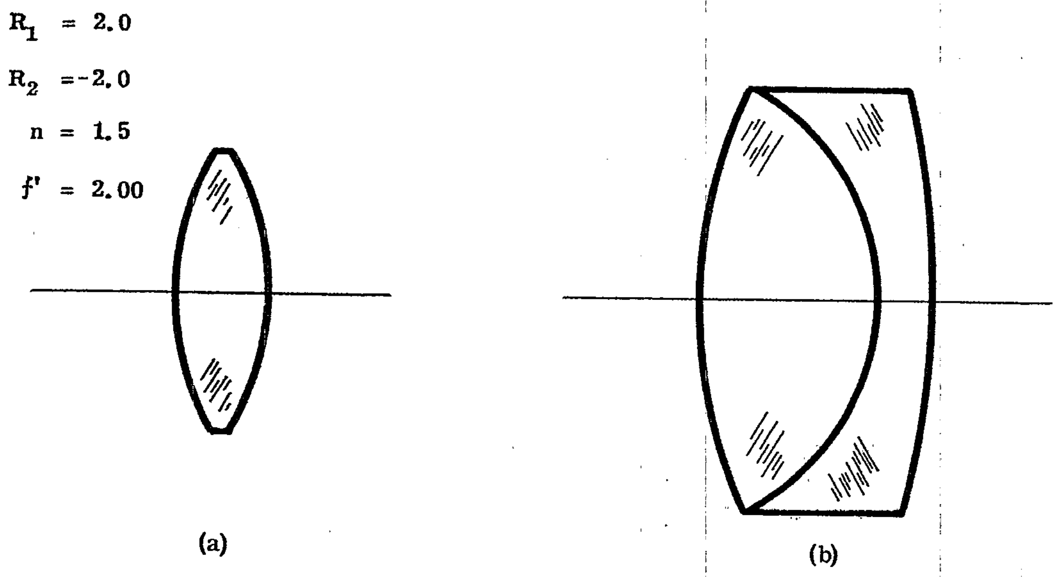
7.3.7 Chromatic aberrations of a simple afocal microscope.

7.3.7.1 Before deciding which of these alternatives is preferable, consider the calculation for axial and lateral color for the system shown in Table 7.2. Since this is an afocal system, the axial beam emerges parallel to the axis, and $u_{k-1} = 0$. Under this condition, it is not possible to use Equations 6-(37) and 6-(38), for the color surface contributions. When $u_{k-1} = 0$ it is necessary to substitute the differential du from Paragraph 6.10.2.2 into Equation 6-(33) and obtain the following equations for the angular chromatic aberrations,

$$du_{k-1} = \alpha \ TAch = \frac{1}{y_k \ n_{k-1}} \sum_{j=1}^{j=k-1} a \qquad (3)$$

and

$$d\bar{u}_{k-1} = \alpha \ Tch = \frac{1}{y_k \ n_{k-1}} \sum_{j=1}^{j=k-1} b \qquad (4)$$



(a)

(b)

Figure 7.3 - (a) is an $f/1$ single lens; (b) is an $f/1.8$ achromatic doublet.

	Object Plane		Entrance Pupil Plane		Lens (a) Objective		Lens(c)		Lens (b)		Exit Pupil Plane
Surface	0		1		2		3		4		5
$-\phi$	0		0		-0.0625		-0.02662		-0.06662		0
t		17.6		0		151		30.02		8.571	
y	0		4.4		4.4		0.625		-0.625		-0.625
u		0.25		0.25		-0.025		-0.04164		0	
$\bar{y}$	1.0		0		0		-8.580		-3.428		0
$\bar{u}$		-0.0568		-0.0568		-0.0568		0.1716		0.400	
ν_{F-C}	∞		∞		60		60		60		
a	0		0		-0.02017		-0.00017		-0.00043		$\Sigma a = -0.0208$
b	0		0		0		0.00238		-0.00238		$\Sigma b = 0$
α TAch											0.0332
α Tch											0

Table 7.4 - Calculations on an afocal microscope with a double lens eyepiece. All lengths are in mm.

These equations are analogous to Equations 6-(37) and 6-(38). The results show that the simple microscope is afflicted with 0.0327 radians of axial color and 0.0076 radians of lateral color. Almost all the axial color is due to the objective. The lateral color is due entirely to the eyelens. The normal observer can detect as little as 0.0003 radians of color fringing assuming that the minimum angle of resolution is about 1' of arc. It is then clear that both the axial and lateral color exceed noticeable amounts of chromatic aberration.

7.3.7.2 The objective lens can be corrected for axial color by making it a doublet. Equations 6-(44) and 6-(45) are used to compute the powers of the separate components. Figure 7.3 (b) shows an $f' = 10$ objective with an f - number of 1.82. It turns out that these curves are again too sharp and the monochromatic aberrations will be difficult to correct. In order to correct the monochromatic aberrations then, it is necessary to flatten the surfaces by dividing the lens into two doublets, each working at $f/3.64$. To do this we divide the entire $|\Delta u| = 0.275$ into two equal parts, each of 0.1375. Each doublet will now work at the same f - number. This value of the f - number $(= 0.5/|\Delta u|)$ will not necessarily equal the value for an infinite object $(= f'/D)$.

7.3.7.3 The chromatic aberration in the eyelens can be corrected in the same way by splitting this lens into two lenses each working at $f/2.18$ and then by achromatizing each part. There is, however, another method which is sometimes used in eyepiece design. A single positive lens may be placed in front of the image plane 3 and adjusted to help refract the chief ray. For such a lens $\bar{y}$ and y will have opposite signs so according to Equation 6-(41) the lens should give a positive lateral color contribution. A lens such as this has been worked out in Table 7.4. The procedure for designing this system was as follows. The extra lens (c) was inserted to the left of the image plane at a position where $y_3 = 0.625$, the same value as the final height of the axial ray but of opposite sign. The chief ray was then traced to the (c) lens intersecting it at $\bar{y} = -8.580$.

7.3.7.4 Since this extra lens is to be used, it should help bend the chief ray. In Table 7.2 the chief ray was bent from -0.0568 to 0.400 by the (b) lens, a total bending of 0.4568. With the (c) lens added, the (b) and (c) lenses should each bend the chief ray by 0.2284. Therefore $\bar{u}_3$ between the (c) and (b) lenses should be 0.1716. This determines ϕ_3, the power of the (c) lens. With ϕ_3 known, u_3 is determined and then t_3 is set so that $y_4 = -0.625$. Thus ϕ_4 is defined. Now the lateral color contribution of a thin lens is proportional to $y\bar{y}\phi/\nu$. $\bar{y}\phi$ is equal to the bending $|\Delta\bar{u}|$ experienced by the chief ray. By making the (b) and (c) lenses refract the chief ray equally and by making $y_3 = -y_4$, the lateral color contributions of the (b) and (c) lenses exactly cancel each other since the ν-values are the same. The axial color of the system is only slightly more under-corrected than the original system in Table 7.2. This chromatic aberration can be eliminated completely by introducing over-correction in the objective lens (a).

7.3.8 Additional effects of adding a field lens.

7.3.8.1 Lens (c) is referred to as a field lens of the eyepiece. (The introduction of a lens near the position of the image due to the objective increases the field of view.) This extra lens (c) has helped the system significantly. The (b) and (c) lenses are $f/2.18$ now and are far more reasonable lenses. There is at this point sufficient reason to expect that this microscope could be corrected to give good imagery. In Section 8 it will be shown that the monochromatic aberrations at an object height of $y_o = 1$ are rather large, so that the final optical design will probably have to have a smaller object field.

7.3.8.2 It should also be noted that the introduction of the (c) lens caused a marked reduction (by a factor of 3) in the distance between the eyelens and the exit pupil. In Table 7.4 this distance is only 8.57 millimeters. This distance, called the eye relief, is too short for comfortable viewing, so some other arrangement of lenses should be found. Without introducing a serious amount of lateral color the (c) lens could be designed with less power. The chief ray would then strike the (b) lens at a larger aperture resulting in an increased distance to the exit pupil. With an eyepiece of this type, the lateral and axial color for the object is fully corrected. However the eyepiece is not color corrected for the plane between the two lenses (b) and (c) where an intermediate image is formed. If cross hairs, or a reticle, is placed in this position it is in effect viewed only by the single eyelens. The reticle will be imaged with lateral color since the single lens is not achromatized. If a reticle is to be used, it is advisable to use an eyepiece that is also color corrected for the intermediate image plane.

7.4 THE TELESCOPE

7.4.1 General. A telescope may be considered as a special case of the microscope, with this slight difference. In the microscope, one compares the visual angle subtended by the image, as viewed through the instrument, with the visual angle subtended by the object at the unaided eye. It is assumed that the observer can place the object at the distance V from the eye. In the telescope it is assumed that the object is inaccessible to the observer. Therefore, in a telescope one compares the visual angles, assuming the observer is always at a fixed distance with respect to the object. This is illustrated in Figure 7.4.

7.4.2 Magnifying power. An object of height $\bar{y}_o$ is located a distance L from an observer. (L is always considered to be positive.) The angle α subtended by the object is $-\bar{y}_o/L$. With the instrument in place, the object is at a distance of z from the first focal point of the objective. A ray from the top of the object, $\bar{y}_o$, passing through the first focal point of the objective strikes the objective at a height $\bar{y}_1 = -\bar{y}_o f_o/z$. This ray then is parallel to the optical axis until it strikes the eyepiece. It then refracts to the second focal point of the eyepiece at an angle with the axis of $\beta = (y_o/z)(f_o/f'_e)$. If the eyepiece is adjusted so that $A = \infty$, or if the eye is located at the second focal point of the eyelens, then β is the apparent angle subtended by the object. Then,

$$MP = -\frac{f_o}{f'_e}\frac{L}{z} = m_o\frac{L}{f'_e} . \qquad (5)$$

This equation actually applies to the microscope by making $L = V$. If L becomes very large as it does for most applications in which telescopes are used, then L/z approaches 1 and, $MP = -f_o/f'_e$. This is the formula usually given for a telescope. At a value of L where L/z is not unity, the MP is increased. Thus it is possible to obtain a magnifying power greater than unity even if $f_o = f'_e$. This makes an interesting optical device. It has unit MP for objects at infinity but greater than unit MP for objects at finite distances.

7.4.3 Objective and eyepiece design. The optics of the objective and eyepiece for the telescope are similar to that of the microscope. The entrance pupil is usually placed at the objective. The eyepiece is usually split into two or more lenses in order to correct for the lateral color. The extra lenses also allow for a wider field of view than one could achieve with a single lens.

7.5 OPTICAL RELAY SYSTEMS. PERISCOPES

7.5.1 Image orientation.

7.5.1.1 In the case of the afocal magnifier, the expression for the MP is V/f'. Since V is always con-

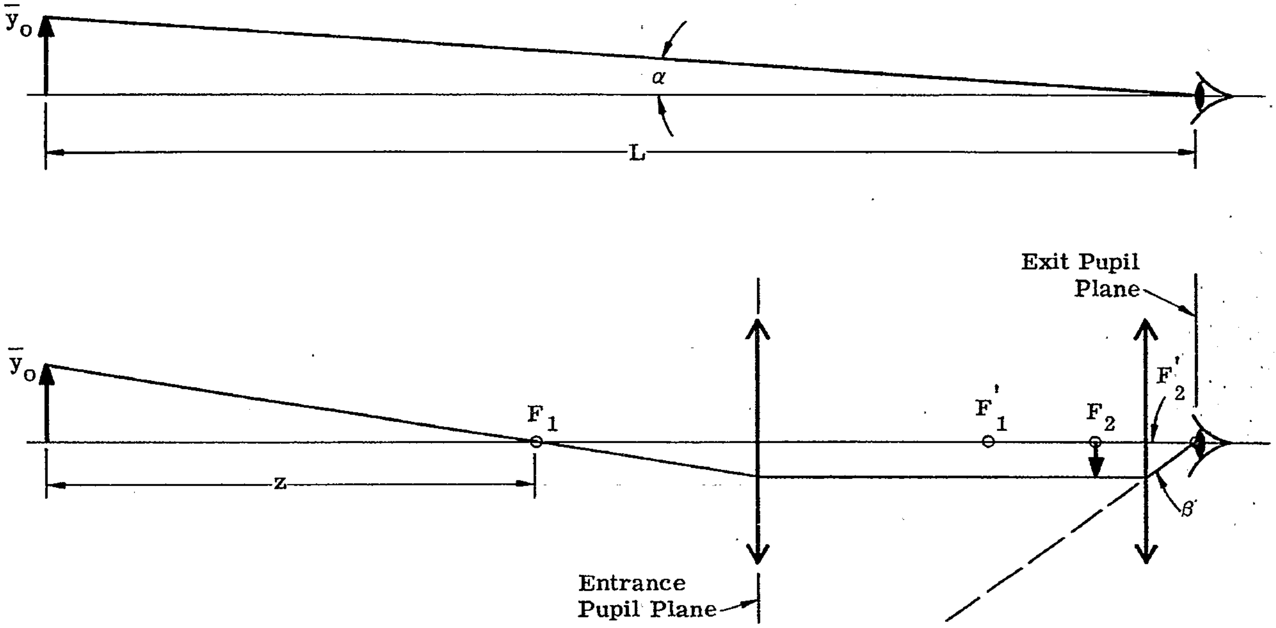
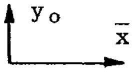
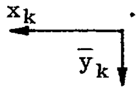


Figure 7.4. The optics of a simple telescope

sidered to be a positive number, the formula indicates that the MP is positive for a positive lens. Positive MP means that the virtual image is in the same orientation as the object.

7.5.1.2 Negative MP or negative magnification means that the image $\bar{y}_k$ is the negative of the object $\bar{y}_o$, in other words the object is inverted; if the optical system is a centered spherical optical system $\bar{x}_k$ will be the negative of $\bar{x}_o$. This means that the object  appears as . Or a letter R will appear as ꓤ . The image is said to be inverted but right handed. This means it is upside down but readable. It appears as a normal R by turning the paper through 180° in its plane.

7.5.1.3 An erect, left-handed image, such as occurs in plane mirrors, would appear as Я . All left-handed images, whether erect or inverted, are unreadable by rotation in the plane only. Left-handed images are sometimes referred to as perverted images. Also see Section 13.

7.5.2 Image inversion for microscopes and telescopes. From Equations (2) and (5) it is seen that a simple microscope and telescope give a negative MP because m_o for an objective is negative. Therefore, these instruments provide a right-handed inverted image. In the microscope it seldom matters if the object is inverted, but in telescopes it is very disturbing to see turned upside down, objects which we are used to seeing erect. Therefore, for telescopes, some means for erecting the image is usually provided. This can be done using prisms or extra lenses. The use of prisms will be described in Section 13. A brief analysis of methods of image erection by lenses will be discussed in the next paragraph.

7.5.3 Image erection by lenses. It is possible to use a second objective in a microscope or telescope to re-image the first image before it is viewed by the eyepiece. This is illustrated in Figure 7.5. The magnifying power for such a system is given by the expression, $MP = m_1 \ m_2 \ L/f'_e$. This procedure can, of course, be carried on with several re-imaging stages if it is desirable to have a long system as in periscopic designs. Since L/f'_e is positive, and each m due to the relaying objectives is negative, it is clear then that if there are an odd number of real images, the MP is negative, while for an even number of real images the MP is positive. A positive overall MP means the image is erect and right-handed.

7.5.4 Field lenses for periscopes. Inspection of Figure 7.5 shows that the size of the object which may be seen through the instrument is definitely limited by the size and permissible f - number of the second

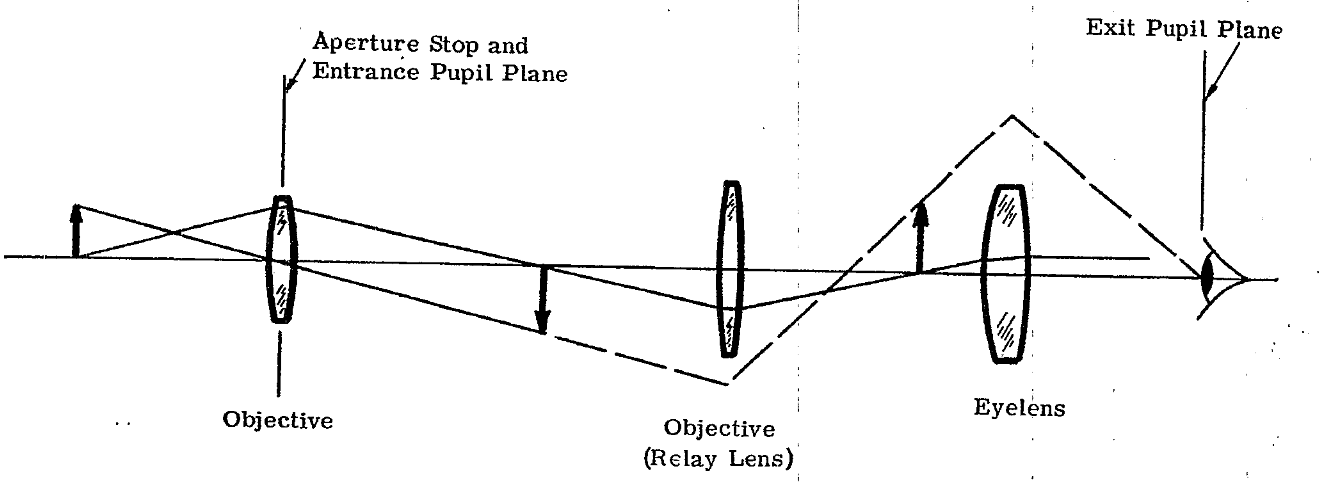


Figure 7.5 - An optical relay system or periscope.

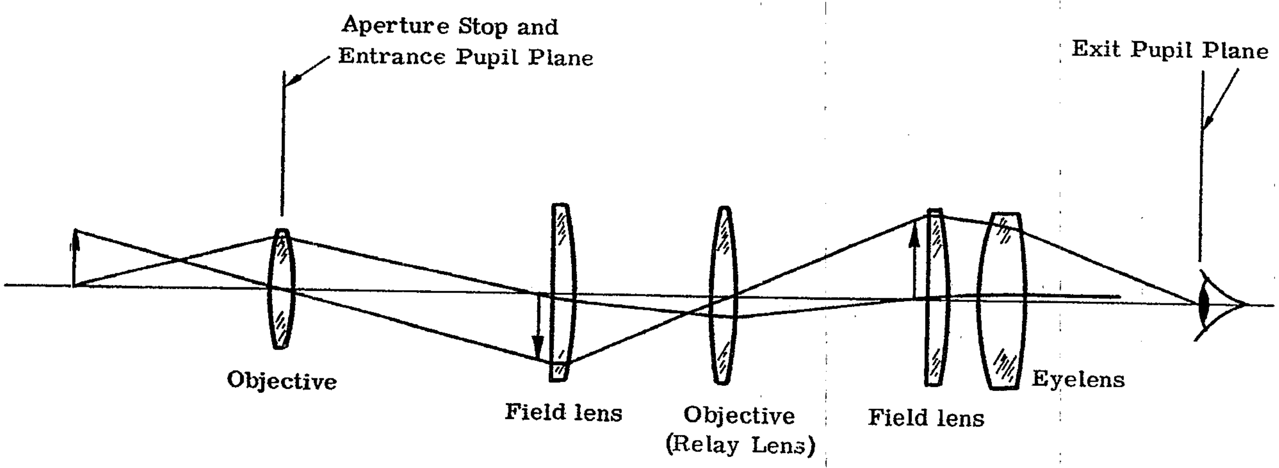


Figure 7.6 - A periscope with field lenses.

objective. The field of view can be increased significantly by introducing extra lenses to help refract the chief ray. This is the same situation described with the eyepiece in Table 7.4. Extra lenses can be introduced at the position of the objective or, if it is desirable to keep the diameter of the system small, the extra lenses can be added near the intermediate images. If the lenses are added near the intermediate images, and therefore referred to as field lenses, they act principally on the chief rays and their primary purpose is to help increase the field of view. Figure 7.6 shows an erecting telescope using field lenses to help increase the field of view.

7.5.5 Position of the aperture stop. The drawing (Figure 7.6) indicates that the first objective is the aperture stop. As drawn, the second objective (relay lens) is larger in diameter than necessary. It could be reduced until the axial ray passed through the margin of the lens, as it does at the first objective. If the diameter of the relay lens were further reduced, this lens would become the aperture stop and the first objective would be too large. In practice the diameters are adjusted so that both objectives are aperture stops.

7.6 THE GALILEAN TELESCOPE

7.6.1 Use of negative eyepiece. In the telescope with $f'_o \gtreqless |f_e|$ it is possible to have a positive or negative focal length eyepiece. If the eyepiece focal length is negative, Equation (5) shows that the MP is positive. The image would therefore be erect. Such a system has very interesting possibilities. A sketch of a telescope of this type is shown in Figure 7.7.

7.6.2 Analysis of the simple Galilean telescope.

7.6.2.1 The exit pupil and aperture stop of this system is usually the pupil of the eye. The entrance pupil is actually located behind the observer's eye, and the size of the objective determines the size of the field of view. The objective is therefore the field stop. A system of this type is worked out in Table 7.5. The table shows the sizes and positions of the entrance and exit pupils. The object field of view (sometimes called the real field) $-\beta\phi_o/\phi_e$ depends on β. In order to obtain an image field of view (sometimes called the apparent field) of β, the $\bar{y}_2$ on the objective lens must be,

$$\bar{y}_2 = \beta \left[MPd - t_2 \right].$$

Therefore, for a given diameter of objective lens, the field of view is determined.

7.6.2.2 For the case of $d = 0$, we have $\bar{y}_2 = -\beta t_2$. Since $f'_o/2\bar{y}_2 = (f\text{ - number})_o$ at which the objective is working for the chief ray,

$$\beta = -\frac{f'_o}{2t_2\,(f\text{ - number})_o}$$

and

$$\alpha = \frac{\beta}{MP} = -\frac{f'_o}{2\,MP\,t_2\,(f\text{ - number})_o}.$$

For MP large compared to unity, the focal length of the objective is large compared to the focal length of the eyepiece. Assuming that $f'_o + f_e = t_2$ can be replaced by f'_o, we have

$$\beta = \frac{1}{2\,(f\text{ - number})_o}$$

and

$$\alpha = \frac{1}{MP\,2\,(f\text{ - number})_o}.$$

These equations show that for a large MP the field of view can be made large only by decreasing $(f\text{ - number})_o$. For example, if $MP = 10$, then $\alpha = 0.05$ radian if the objective is $f/1$. An $f/1$ lens is very difficult to make. The usual doublet achromat would have only an $f/3$ aperture. For such a doublet objective $\alpha = 0.017$ radian or 0.95°. Then $\beta = 9.5°$, which is a very small apparent field of view.

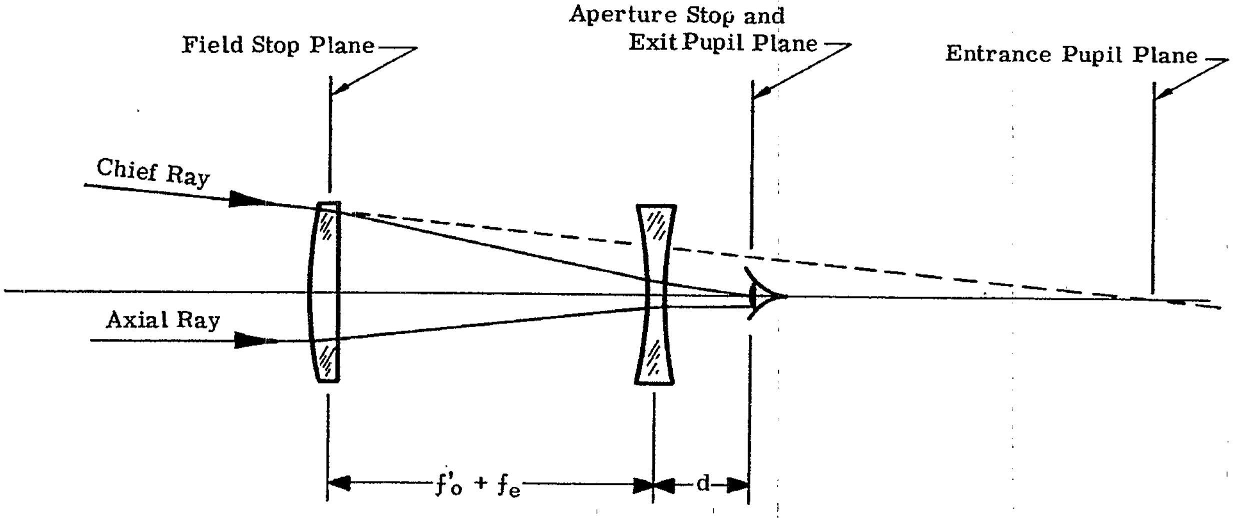


Figure 7.7 - The Galilean telescope

Surface	Object 0		Entrance Pupil 1		Objective 2		Eyepiece 3		Exit Pupil 4
$-\phi$	0		0		$-\phi_o$		$-\phi_e$		0
t		∞		$-MP\left[dMP - t_2\right]$		$(f'_o + f_e)$		d	
y	0		$-r\frac{\phi_e}{\phi_o}$		$-r\frac{\phi_e}{\phi_o}$		r		r
u		0		0		$r\phi_e$		0	
$\bar{y}$	∞		0		$\beta\left[dMP - t_2\right]$		$-d\beta$		0
$\bar{u}$		$-\beta\frac{\phi_o}{\phi_e}$		$-\beta\frac{\phi_o}{\phi_e}$		$\beta - \phi_e d\beta$		β	

Table 7.5 - Calculations for a Galilean telescope.

8 ABERRATION ANALYSIS AND THIRD ORDER THEORY

8.1 SIGNIFICANCE OF RAY TRACE DATA

8.1.1 First order system.

8.1.1.1 In Sections 5, 6, and 7, formulae and techniques were presented to enable the designer to set up a first order optical system. As an aid in arriving at a final first order solution, it is customary to trace two paraxial rays. One of these starts from an object point on the optical axis and heads toward the tentative edge of the entrance pupil. The other ray starts from an object point at the tentative edge of the field and heads toward the tentative center of the entrance pupil.

8.1.1.2 The data from these two rays may be used to determine the trace of any other paraxial ray through the system. The magnification, focal length, and chromatic aberration may be calculated. The planes of the paraxial image, entrance pupil, aperture stop, exit pupil, and field stop may be finally located, and the sizes of the pupils and stops can be finally determined. The f - number and fields of view can be calculated.

8.1.1.3 The calculation of the above characteristics of a first order solution has already been discussed. Additional calculations using paraxial ray trace data will be given later in this section where the aberrations of a system will be analyzed, and a third order theory will be developed which will provide understanding of the sources of image errors and suggest methods for correcting these errors.

8.1.2 Skew ray trace. After a first order system has been set up, skew and meridional rays are traced by the methods discussed in Section 5. This tracing of skew rays provides the basic method of investigating lens performance. The paraxial image and the entrance pupil furnish excellent reference planes which are used in the interpretation of non-paraxial ray trace data.

8.2 THE SPOT DIAGRAM

8.2.1 Representation of ray trace data. One way to make a graphical summation of ray trace data is by means of a spot diagram. Such a diagram is a plot of the intersection coordinates in the reference planes of rays traced through the lens or system from a single object point. The two reference planes usually chosen are the entrance pupil plane and the paraxial image plane. The rays traced from the object point are usually chosen so as to form a uniform pattern of intersection with the entrance pupil plane while the resulting image is represented by the ray intersections in the paraxial image plane. Figures 8.1 (a) and 8.1 (b) show typical spot diagrams of this type for rays traced from an object point (object distance not specified) which lies in the YZ (meridional) plane at coordinates $X_o = 0$ and $Y_o =$ an arbitrary value. Thus, the twelve spots at $X_1 = 0$ in Figure 8.1 (a) represent meridional rays; all others are skew rays. In Figure 8.1 (b), the meridional rays are at $X_k = 0$. There is a one to one correspondence between the spots in the two diagrams. In general, the spots at large values of X_1 correspond to the spots at large X_k . The Y_k axis is an axis of symmetry because the Y_1 axis is an axis of symmetry.

8.2.2 Ray distribution in the entrance pupil. The shape of the entrance pupil may be found with sufficient accuracy for most applications from paraxial ray tracing by the method described in Section 6. With automatic high speed computers it is possible to trace a regular grid of rays through the system. If any ray does not pass through every clear aperture the ray is rejected. With a computer program of this type, the shape of the vignetted aperture is automatically found as the boundary of the non-rejected rays.

8.2.3 Ray distribution in the image plane.

8.2.3.1 The spot diagram shown in Figure 8.1 (b) is extremely useful to a designer in evaluating a system. The diagram indicates how well the lens concentrates the energy from the object point into an image point. One can count the number of points in concentric circles in the image plane and obtain what is called an energy distribution curve. In Figure 8.1 (a) there are 192 points in the entrance pupil. If it is assumed that each point represents an equal amount of energy, a given point is equivalent to 1/192 of the total energy from the object point passing through the aperture. Now by drawing concentric circles around what appears to be the center of concentration of spots, and counting the number of spots within each circle one obtains the total energy as a function of (circular) image size. Figure 8.2 is a plot of percent energy versus image size for the spot diagram shown in Figure 8.1 (b). In a theoretically perfect geometrical image all the spots would be concentrated at a point. However, in the case of the image due to a perfect optical system, the geometrical image is only an approximation; the actual image formed would be larger than a point due to diffraction effects.

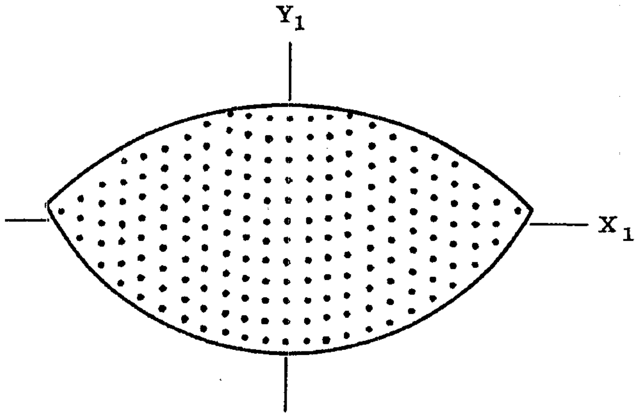


Figure 8.1 (a)- Spot diagram of rays passing through the entrance pupil.

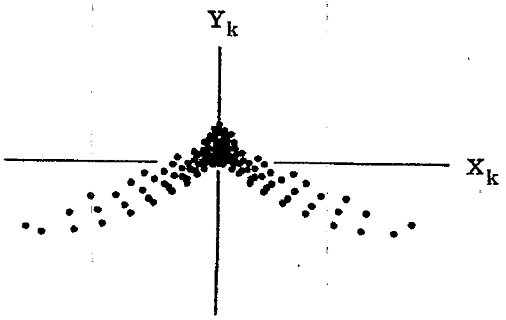


Figure 8.1 (b) - Spot diagram of rays incident on the paraxial image plane.

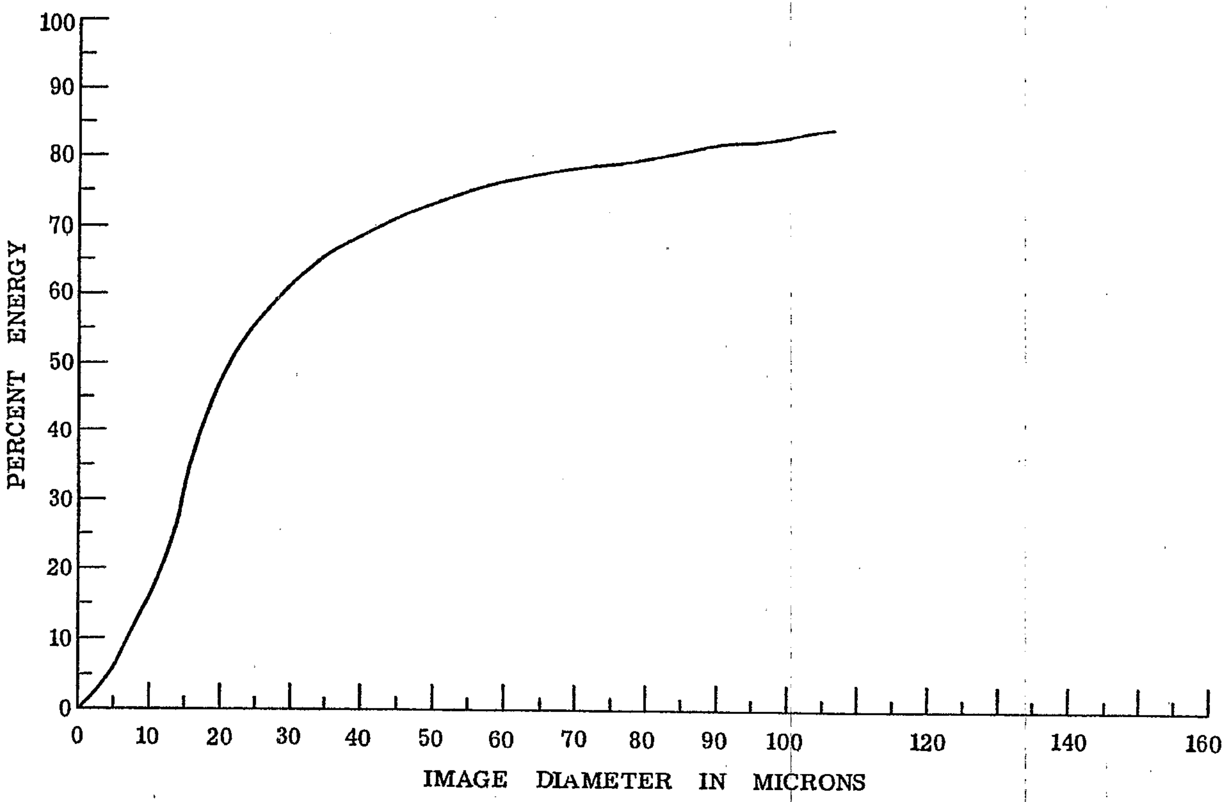


Figure 8.2- Energy distribution for the image shown in Figure 8.1 (b).

8.2.3.2 The spot diagram is useful in evaluating the image performance of a lens but it gives little insight to a designer as to why an image is spread out. In making a spot diagram, no attempt is usually made to identify each ray; therefore the designer usually has no means of visualizing what happens to the ray as it passes through the lens. Of course, it is perfectly possible to program the computer so that each ray in Figure 8.1 (b) is identified with a ray in Figure 8.1 (a), but with the large number of rays usually chosen for energy distribution representation, this would be unnecessarily complex. Instead, in order to understand the reason for image deformation, it is common practice to trace only a few selected rays through the aperture and plot the data in a different manner.

8.3 MERIDIONAL AND SKEW FANS

8.3.1 General method.

8.3.1.1 Instead of using spot diagrams and energy distribution functions, the ray trace data usually may be more conveniently analyzed by the method of meridional and skew ray fans. In using this method a common practice is to trace from a selected object point in the YZ plane, five to seven meridional rays (rays lying in the YZ plane) through the system. These rays, called the meridional ray fan, are chosen to intersect the vignetted entrance pupil in a nearly uniform spread, with upper and lower extremes (the rim rays) as close as possible to the respective vignetted pupil limits. See Figure 8.3. One of these rays is chosen to intersect the entrance pupil at $X_1 = 0$, $Y_1 = 0$ and thus, by definition, is the chief ray. The angle between the chief ray from the given object point and the optical axis is used to identify the meridional ray fan and its associated skew ray fan. The latter is constructed by tracing three to five rays from the same object point, which enter the vignetted entrance pupil at the intersection of the pupil and the XZ plane, i.e., at $Y_1 = 0$. Rays having positive X values only are needed since the object point lies in the meridional plane and the system is therefore symmetrical about the YZ plane. On the other hand, since the object point is not necessarily on the optical axis, rays above and below the Z axis are not symmetrical, so that rays must be traced having both positive and negative Y values at the entrance pupil plane.

8.3.1.2 Since they lie in a plane throughout their passage through the system, the behavior of the meridional rays can be well understood by making a plot of the coordinates of each ray intercept in the image plane (Y_k) versus the corresponding ray intercept in the entrance pupil plane (Y_1). This, in effect, is a similar but much more accurate presentation of the ray height data which could be obtained through graphical ray tracing.

8.3.1.3 Skew rays, on the other hand, do not usually remain in a single plane during their passage through the system. Thus, even though we have simplified the problem by choosing only those that intersect the entrance pupil plane on the X_1 axis ($Y_1 = 0$), they will normally have both X and Y coordinates in the image plane. Thus, for skew rays, it is necessary to make two types of plots: X_k versus X_1, and Y_k versus X_1. For perfect geometrical imagery, these plots would be straight lines of zero slope.

8.3.2 Illustrative example. In the following paragraphs, the arrangement and interpretation of these three curves will be discussed in detail. The example to be used will employ the same lens as shown in Table 6.7, except that in the table, the entrance pupil plane was not included, therefore surface 1 is the first lens surface. However, in the following discussions, surface 1 will be the entrance pupil plane, surface 2 the first lens surface, and so on. This is illustrated in Figure 8.4, which is drawn to scale from Table 6.7. The lens is a typical photographic Taylor triplet. The object surface for the lens is at infinity; the entrance pupil plane is located 2.2 cms to the right of the first surface of the lens. Rays representing fans of obliquities of 0°, 10°, 15°, and 20° have been traced into the system. (Note: with the object at infinity, the term "fan" is somewhat of a misnomer since all rays from a given object point are parallel, a situation which would not exist if the object were at a finite distance). The diagram shows the path of the extreme upper and lower rays for field angles of 10° and 20°. The upper rim ray at 0° is also shown; the lower is similar by symmetry. Notice how the upper and lower rays at 10° and 20° do not pass through the edge of the aperture stop. This is because the designer decided to vignette the oblique rays in order to eliminate some badly aberrated rays. The back focal length (BFL) is the distance between the last surface of the last element (surface 7) and the second focal point. Table 8.1 gives the numerical values for this lens.

8.4 USE OF THIRD ORDER THEORY IN ABERRATION ANALYSIS

8.4.1 Ray trace data. The numerical data used in the following discussions are the results of paraxial, meridional and skew ray traces for the lens shown in Table 8.1 and Figure 8.4. This lens, with very slightly different ν - number, was given in Tables 6.6 and 6.7.

8.4.2 Analysis of data. The curves of ray trace data will be plotted and analyzed in a manner that will be helpful to the designer trying to minimize the aberrations. The plots and analyses will make use of the third order theory to investigate the third order aberrations which, as explained in Section 5.11.3 are the first

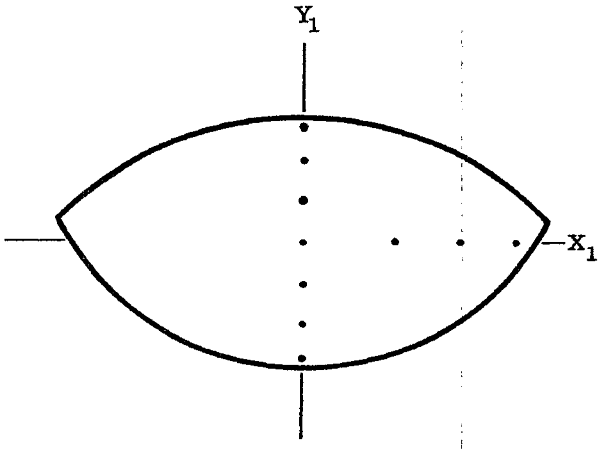


Figure 8.3 - Positions in the entrance pupil of selected meridional rays and skew rays used to analyze images.

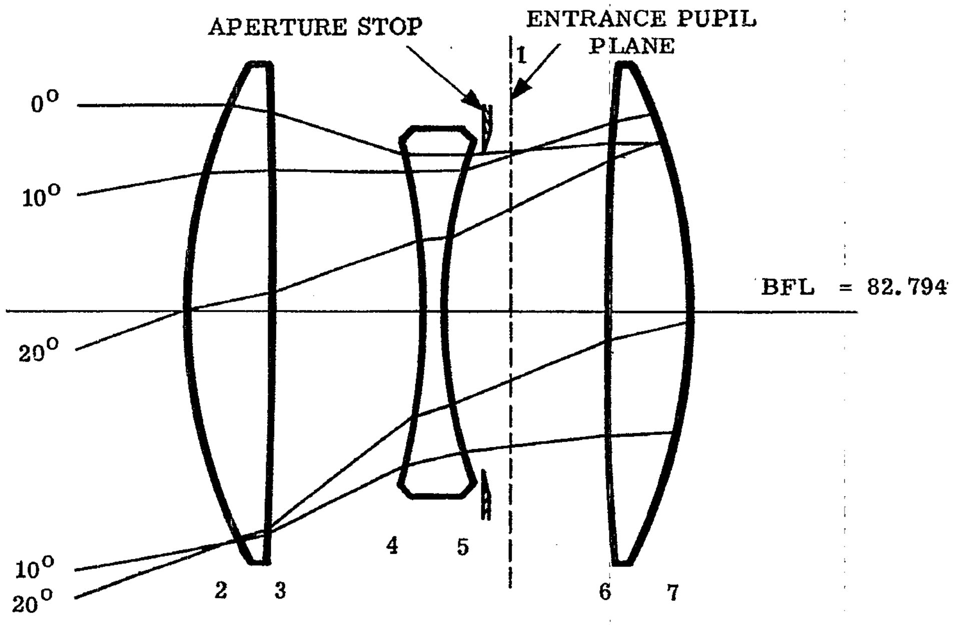


Figure 8.4 - Sample lens used for numerical analysis.

Surface	Radius	Thickness	n_D	ν-no.
2	39.55			
		6.0	1.620	60.3
3	-678.43			
		10.654	1.0	
4	-50.15			
		1.5	1.621	36.2
5	38.50			
		11.369	1.0	
6	197.43			
		6.0	1.620	60.3
7	-40.67			

Table 8.1 - Numerical values for lens in Figure 8.4. All lengths in millimeters. The numerical values are exact except for the radii. Exact curvatures are given in Table 8.2.

approximations to the aberrations. The method of plotting differs from that described in Section 8.3 in that only the essential information is shown. That is the difference in the paraxial plane intercepts for the chief ray and the other rays of the fans is plotted since it is this difference, or lack of coincidence, which the designer is trying to overcome.

8.5 THE 0° IMAGE IN D LIGHT

8.5.1 The 0° image polynomial.

8.5.1.1 The image of an axial object point at infinity is studied by tracing three meridional rays with $K_o = 0$, $L_o = 0$, $M_o = 1$ at values of $Y_1 = 1.5$, 1.0 and 0.5. For meridional rays from an axial object point, negative values of Y_j are symmetrical with the positive values. The results are plotted and encircled in Figure 8.5. The vertical scale is labelled $(Y_k - \overline{Y}_k)$. $\overline{Y}_k$ is the height of the chief ray $(\overline{Y}_1 = 0)$ on the final paraxial $(y_k = 0)$ image plane. For these axial rays $\overline{Y}_k = 0$. The circled points, connected by the full curve, can be fitted fairly accurately to a power series of the form

$$Y_k - \overline{Y}_k = b_1 Y_1 + b_3 Y_1^3 + b_5 Y_1^5 + O(7). \qquad (1)$$

The letters b_3, b_5, etc. are called the spherical aberration coefficients. The term $O(7)$ stands for all the terms of order 7 and above, as explained in Paragraph 5.5.2.3.

8.5.1.2 When ray data are plotted in this manner the slope of a line drawn between any two ray points on the curve is proportional to the longitudinal distance from the paraxial image plane to the plane where the two rays focus. That this is true, can be seen from Figure 8.6. This diagram shows two actual rays converging towards the image surface. The image surface, where the two rays focus, will be called the $k + 1$ surface. The paraxial image plane is called the kth surface. By placing the paraxial image plane to the left of the intersections of the optical axis with rays (a) and (b), the two Y_k values are positive. Such a diagram would not represent the physical situation of a single converging lens, because for that case, the paraxial image plane is to the right of the intersection points of the optical axis with non-paraxial rays. The situation represented in Figure 8.6 could be attained, for example, by the forming of an image by a diverging system of an unaberrated, virtual object.

8.5.1.3 From the diagram we have

$$Y_{ka} = Y_{(k+1)a} - t_k \tan U_{(k-1)a}$$

and

$$Y_{kb} = Y_{(k+1)b} - t_k \tan U_{(k-1)b}.$$

Since $Y_{(k+1)a} = Y_{(k+1)b}$, subtraction gives

$$t_k = -\frac{Y_{ka} - Y_{kb}}{\tan U_{(k-1)a} - \tan U_{(k-1)b}}.$$

This equation and Figure 8.6 apply to two non-paraxial rays. It will be assumed that the following relation is a valid approximation for either of these rays; namely

$$\frac{Y_1}{\tan U_{k-1}} = \frac{y_1}{u_{k-1}}.$$

If these rays were paraxial, this relation would be exact; assuming it to hold approximately for non-paraxial rays, there results

$$\tan U_{(k-1)a} - \tan U_{(k-1)b} = \frac{Y_{1a}}{y_1 / u_{k-1}} - \frac{Y_{1b}}{y_1 / u_{k-1}},$$

and finally,

$$t_k = -\frac{y_1}{u_{k-1}} \left(\frac{Y_{kb} - Y_{ka}}{Y_{1b} - Y_{1a}} \right). \qquad (2)$$

When the object point is at infinity, then $-y_1 / u_{k-1} = f'$. Equation (2) is only an approximation for non-paraxial rays. But at worst it gives the order of magnitude of t_k; this is all that is needed for

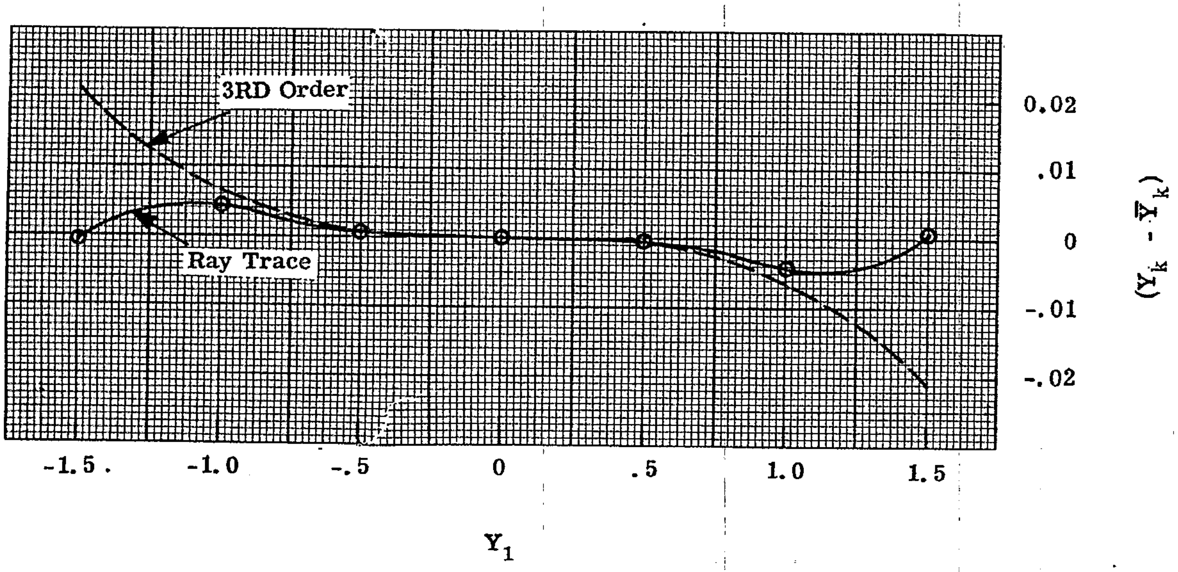


Figure 8.5 - Meridional ray plot at 0°.

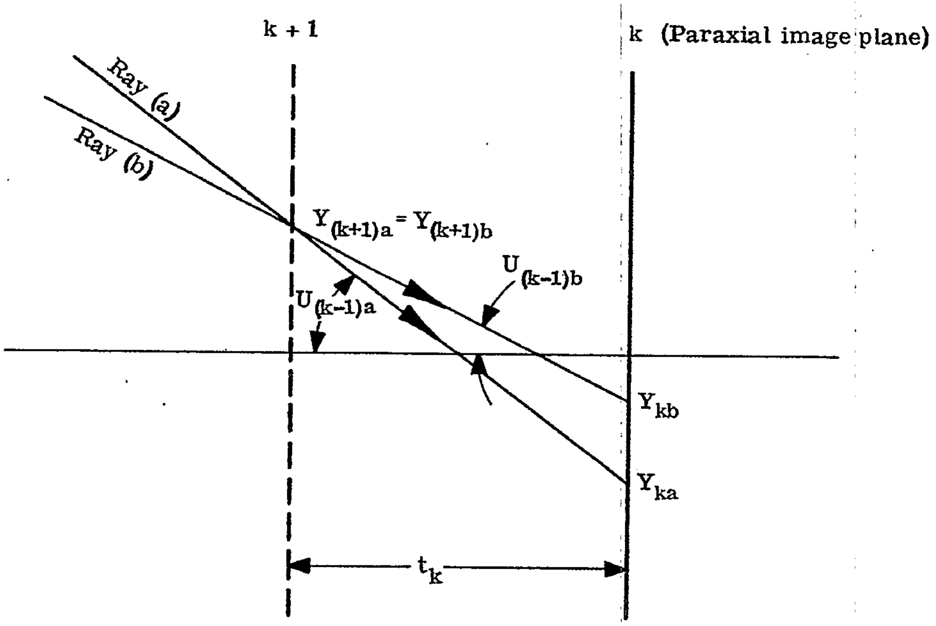


Figure 8.6 - Diagram illustrating shift of focus relationships.

third order design procedure.

8.5.1.4 Since the slope of the line connecting any two points on the $(Y_k - \overline{Y}_k)$ versus Y_1 curve is $\Delta Y_k / \Delta Y_1$, this slope is proportional to the distance from the paraxial image plane to the plane of focus for the two rays. As the two rays approach each other, the two points on the curve do likewise, and the slope of the chord approaches the slope of the curve. Hence the slope of the curve at any point is proportional to the distance from the paraxial image plane to the plane of focus for infinitely close rays. From ray trace data resulting in Figure 8.5, and from paraxial ray trace data giving y_1 / u_{k-1}, we are able to determine the shift (t_k) of the image plane from the paraxial image plane.

8.5.1.5 Now since the ray data shown in Figure 8.5 was obtained in the paraxial image plane the slope of the curve must be zero at $Y_1 = 0$. Therefore b_1 will be zero in Equation (1). The presence of a linear term indicated by a slope different from zero at $Y_1 = 0$, means that the paraxial rays are not focused in the final image plane upon which the ray heights are calculated. This linear term in Equation (1) can be eliminated by shifting the plane upon which the ray heights are calculated. When this has been done, so that the slope of the curve is zero at $Y_1 = 0$, and the linear term is absent, any further deviation, $(Y_k - \overline{Y}_k) \neq 0$, indicates the presence of spherical aberration. Therefore, the first approximation to the spherical aberration, written as $(Y_k - \overline{Y}_k)$, varies as the cube of the entrance pupil radius. This part of the spherical aberration, the third order spherical aberration, would vary with Y_1 as shown by the dashed line in Figure 8.5.

8.5.1.6 Because the slope of the line between any two points on the curve is proportional to the t_k for the two rays considered, Figure 8.5 shows that the rays traced at $Y_1 = 0.5$ and 1.0 are focused closer to the lens than the paraxial image plane; while the ray at $Y_1 = 1.5$ is focused almost exactly on the paraxial image plane. In this system, the third order coefficient, b_3, is negative and the lens is said to be undercorrected for the third order spherical aberration. It is called undercorrected spherical aberration because a single positive lens has spherical aberration of this sign. (See Paragraph 6.10.5.1). The coefficient, b_5, is called the fifth order coefficient. In this case it is positive because the full curve (Figure 8.5) is between the third order curve and $Y_k - \overline{Y}_k = 0$; hence the fifth order term [Equation(1)] has a sign opposite to that of the third order term. The fifth order coefficient is said to be overcorrected.

8.5.2 The third order aberration coefficient.

8.5.2.1 Now a truly remarkable feature of optical systems is that the coefficient, b_3, may be computed from axial paraxial ray data. This is done by calculating B_j, the third order spherical aberration surface contribution, at each surface in the optical system. Then,

$$b_3 = -\left[\frac{1}{2\,(n_{k-1}\;u_{(k-1)})\,y_1^3}\right]\sum_{j=1}^{j=k-1} B_j \;, *$$

where y_1 is the height of the axial paraxial ray in the entrance pupil plane and u_{k-1} is the final angle with the optical axis for this ray. Therefore, since $\overline{Y}_k = 0$ for 0° obliquity, the third order approximation for Y_k is,

$${}_3Y_k = -\frac{\Sigma B}{2\,(n_{k-1}\;u_{k-1})}\left(\frac{Y_1}{y_1}\right)^3 . ** \qquad (3)$$

8.5.2.2 B_j is calculated from the axial paraxial ray data for each surface with the following formulae

$$B = S\,i^2 , \qquad (4)$$

and

$$S = y\,n_{-1}\left(\frac{n_{-1}}{n} - 1\right)(u + i) . \qquad (5)$$

In Table 8.2, B_j is calculated for each surface of the sample lens being studied. This is the same lens

---

* Up to this point in the text an attempt has been made to derive the equations, or to indicate specifically how they may be derived. This practice will no longer be followed; thus, equations may be presented without proof. To do otherwise would necessitate lengthy and complex departures from the main train of thought.

** In later sections the symbol Σ will be used to indicate the summation of all surface contributions. The proper summation limits will be eliminated.

Surface	Object 0	Entrance Pupil 1	2	3	4	5	6	7	Image 8
c		0	0.25285	-0.01474	-0.19942	0.25973	0.05065	-0.24588	0
t		-2.20000	0.60000	1.06541	0.15000	1.13691	0.60000	8.27937	
n	1.0000	1.0000(	1.62000	1.00000	1.62100	1.00000	1.62000	1.00000	
$(n_{-1}/n)-1$			-0.382716	0.62000	-0.383097	0.62100	-0.382716	0.62000	
y		1.50000	1.50000	1.41291	1.14862	1.13883	1.22735	1.24192	0
u	0	0	-0.145155	-0.24806	-0.065279	0.077866	0.02427	-0.15000	
i		0	0.37928	-0.16598	-0.47712	0.230508	0.14003	-0.281089	
u + i		0	0.23412	-0.41404	-0.54239	0.30837	0.16431	-0.43109	
B			-0.01933	-0.01619	0.05433	0.01878	-0.00151	-0.04249	$\Sigma B = -0.00641$
$\bar{y}$		0	-0.80073	-0.61944	0.09189	-0.04712	0.49425	0.66486	$_3Y_k = -0.02135$
$\bar{u}$	0.36397	0.36397	0.30216	0.49516	0.29845	0.47618	0.28436	0.35930	3.63966
$\bar{i}$		0.36397	0.16150	0.31129	0.51348	0.28621	0.50121	0.12088	
F			-0.00823	0.03036	-0.05847	0.02332	-0.00542	0.01827	$\Sigma F = -0.000170$
C			-0.00351	-0.05694	0.06293	0.02896	-0.01939	-0.00786	$\Sigma C = -0.004200$
E			-0.01378	0.10994	-0.09223	0.07278	-0.09008	0.01544	$\Sigma E = -0.002070$
P			-0.09677	-0.00564	0.07640	0.09950	-0.01938	-0.09410	$\Sigma P = -0.04000$
dn/n		0	0.00635	0	0.01058	0	0.00635	0	
$\Delta(dn/n)$			0.00635	-0.00635	0.01058	-0.01058	0.00635	-0.00635	
a			-0.00362	-0.00242	0.00580	0.00450	-0.00109	-0.00361	$\Sigma a = -0.00040$
b			-0.00154	0.00452	-0.00624	0.00559	-0.00390	0.00154	$\Sigma b = -0.000026$

Table 8.2 - Third order calculations on triplet lens in Figure 8.4.

illustrated in Table 6.7. The dotted curve in Figure 8.5 shows the third order curve as predicted by Equation (3). One notes that at $Y_1 = 1.5$, the dotted curve passes through the point $Y_k = -0.0214$. Notice also how the third order curve follows the true aberration curve very closely out to $Y_1 = 0.75$.

8.5.2.3 Returning to Equation (1) it follows that $b_3 = -0.0214/y_1^3 = -0.006341$. Since the actual ray traced at $Y_1 = 1.5$ strikes the final image plane at $Y_k = 0$ it is possible to compute that $b_5 = 0.002818$ if it is assumed that $O(7) = 0$. By using Equation (1) then at $Y_1 = 1.0$, Y_k should equal -0.00352. The actual ray traced point comes out at $Y_k = -0.0041$. This difference, 0.0006, is small compared to the total spherical aberration -0.0041. This means that the spherical aberration curve shown in Figure 8.5 may be approximately obtained by calculating the third order coefficient b_3 from axial paraxial ray data, and tracing one non-paraxial ray. On the other hand, the curve can also be obtained by tracing two non-paraxial rays, and calculating b_3 and b_5. Since the third order coefficient calculation depends on the individual surface contributions, it helps the designer see the source of the aberrations. For the example shown in Table 8.2, we see that the first two, and the last two surfaces of the lens give negative spherical aberration. The two surfaces of the central negative element provide all the positive or over-correction. The contribution on surface number four of the lens has the largest positive value. This coupled with the large angle of incidence on this surface is the main reason that the fifth order coefficient is positive. If one wanted to reduce the fifth order coefficient, it would be necessary to find a solution with a smaller angle of incidence on this surface or a smaller spherical aberration coefficient. If the fifth order coefficient were reduced, the total aberration (full curve) will be closer to the third order. The maximum under-correction, which now occurs at about $Y_1 = 1.1$, would increase and would occur at a larger Y_1. Such a lens would exhibit an increased zonal spherical aberration. The point of zero aberration, now at $Y_1 = 1.5$, would increase towards larger values of Y_1, so that the lens could be used at a larger aperture.

8.5.2.4 The Y_k versus Y_1 curves shown in Figure 8.5 were obtained in D light. Similar calculations could be made in F and C light. The value of b_3 can vary with wavelength, and since the plot is made for the paraxial focal plane in D light, the F and C paraxial rays will focus farther from the lens by approximately $f'/2200$. Therefore b_1 for F and C light, if we have a true (F - C) achromat, will be positive and equal to 1/2200. On this scale this is a negligible amount of aberration amounting to one-tenth of the zonal aberration, 0.0041. The F and C curves, corresponding to Figure 8.5, would have a positive slope at $Y_1 = 0$.

8.5.3 The Seidel spherical aberration.

8.5.3.1 Equations (3), (4) and (5) give the calculation of ${}_3Y_k$, the third order approximation to Y_k. Because $\overline{Y}_k = 0$, and hence for an unaberrated image point $Y_k = 0$, ${}_3Y_k$ is the third order approximation to the spherical aberration, measured in a plane perpendicular to the optical axis. Hence, it is sometimes referred to as the transverse spherical aberration. In the following section the aberrations of an off-axis image point will also be expressed as transverse aberrations.

8.5.3.2 Another measure of spherical aberration, called the longitudinal spherical aberration, is the distance along the optical axis between the paraxial image plane and the non-paraxial ray. The third order approximation to the longitudinal spherical aberration, referred to as the Seidel longitudinal spherical aberration, is numerically equal to ${}_3Y_k / u_{k-1}$. Hence, from an expression for the Seidel aberration, Equations (3), (4) and (5) readily follow.

8.6 IMAGERY FOR AN OFF-AXIS OBJECT POINT

8.6.1 The oblique image polynomial.

8.6.1.1 The solid curve in Figure 8.7 (a) is a plot of meridional rays traced through the sample lens at 10° $(K_o = 0, L_o = 0.1736)$. The coordinates for the entering rays on the entrance pupil extend from $Y_1 = 1.35$ to $Y_1 = -1.35$. The vertical scale is again $(Y_k - \overline{Y}_k)$. The curve represents the displacement between the ray heights and the chief ray height in the paraxial image plane. This curve may also be represented by a power series. The power series can be expressed in different ways, but the following uses the well known Seidel third order coefficients. The polynomial can be expressed for any ray coordinate (Y_1, X_1) in the entrance pupil for any object height $\overline{Y}_o$ $(\overline{X}_o = 0)$. Hence the series are sufficiently general so that they can be used with skew rays. There are two equations, one for $(Y_k - \overline{Y}_k)$, and the other for X_k . $\bar{X}_k$ is always zero. These equations are

$$Y_k - \bar{Y}_k = -\frac{1}{2(n_{k-1}u_{k-1})}\left[\Sigma B(Y_1^2 + X_1^2)\frac{Y_1}{y_1^3} + \Sigma F\frac{3Y_1^2 + X_1^2}{y_1^2}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right) + \Sigma(3C + P\Phi^2)\frac{Y_1}{y_1}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2\right] + O(5), \quad (6)$$

and

$$X_k = -\frac{1}{2(n_{k-1}u_{k-1})}\left[\Sigma B(Y_1^2 + X_1^2)\frac{X_1}{y_1^3} + \Sigma F\frac{(2Y_1X_1)}{y_1^2}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right) + \Sigma(C + P\Phi^2)\left(\frac{X_1}{y_1}\right)\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2\right] + O(5). \quad (7)$$

8.6.1.2 The expressions ΣB , ΣF , ΣC and ΣP are the sums of the third order surface contributions for spherical aberration, coma, astigmatism and Petzval curvature. (C must not be confused with c , the curvature of a surface.) The terms, y_1 , $\bar{y}_o$, and $n_{k-1}\ u_{k-1}$, are the data from the two paraxial rays traced through the system. $\bar{Y}_o$ is the object point height. Y_1 and X_1 are coordinates for a general ray in the entrance pupil. If the object point is at infinity, as it is in the example being described, then $\bar{Y}_o / \bar{y}_o$ should be replaced by $(\tan \bar{U}_o)/\bar{u}_o$ or $L_o / M_o\ \bar{u}_o$.

8.6.1.3 If ΣB , ΣF , ΣC and ΣP are known, Equations (6) and (7) can be used to predict the position coordinates of any ray in the image surface corresponding to a given point in the object surface. The accuracy of the prediction depends on the magnitude of aberrations higher than third order. According to the first order theory, the chief ray should strike the image plane at $\bar{Y}_k = \bar{Y}_o\ m$ if $\bar{Y}_o$ is finite, or at $\bar{Y}_k = f \tan \bar{U}_o$, if the object is at infinity. However, the actual chief ray is displaced from the ideal image point due to a fifth aberration, distortion. There is also a polynomial to express this displacement.

$$(\bar{Y}_k - \bar{Y}_o\ m) = -\frac{\Sigma E}{2(n_{k-1}\ u_{k-1})}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^3 + O(5), \quad (8)$$

where ΣE is the third order contribution for distortion. Equation (8) can be included in Equation (6) but it was not because it is somewhat easier to plot $(Y_k - \bar{Y}_k)$ as has been done in Figure 8.7. The fractional distortion which is defined by the ratio $(\bar{Y}_k - \bar{Y}_o\ m)/\bar{Y}_o\ m$ may be written to read as follows,

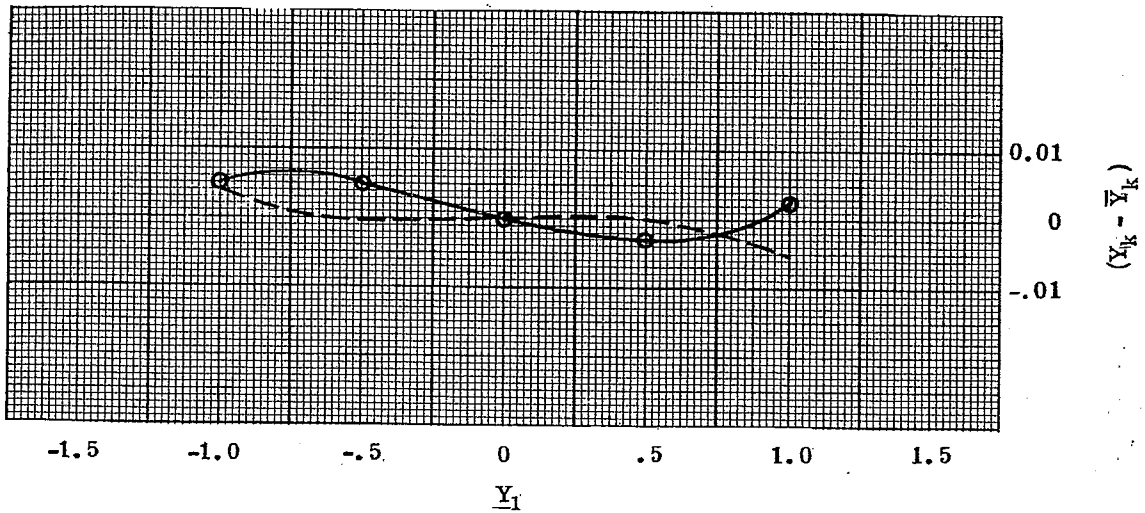
$$\text{fractional distortion} = \frac{\bar{Y}_k - \bar{Y}_o\ m}{\bar{Y}_o\ m} \doteq -\frac{\Sigma E}{2\Phi}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2.$$

Note that the fractional distortion varies with the square of the object height ratio $(\bar{Y}_o/\bar{y}_o)$. In Section 8.7 the method for calculating B , F , C , E and P will be described. The actual calculations for the sample problem are shown in Table 8.2.

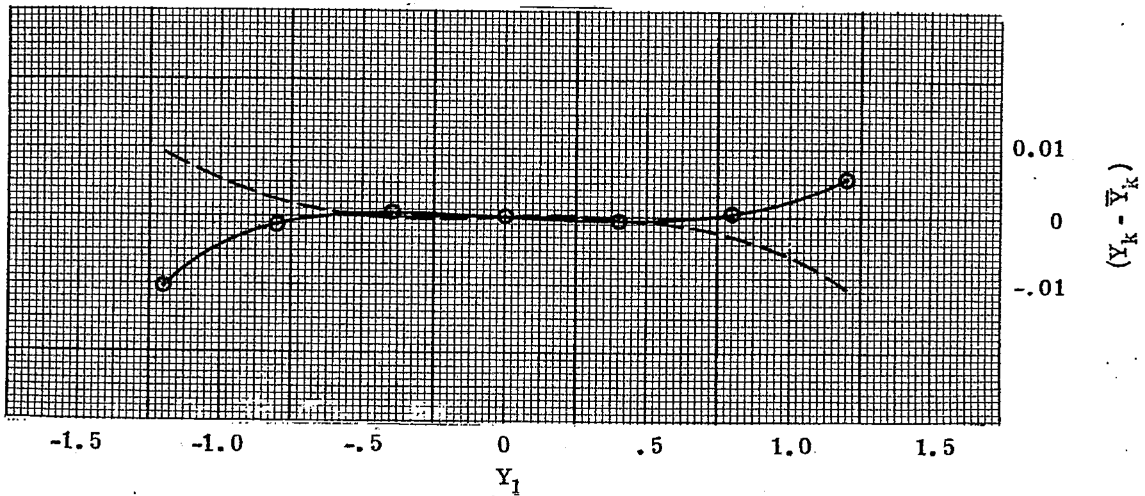
8.6.2 Examples of third order aberrations.

8.6.2.1 The third order ray predictions for $(Y_k - \bar{Y}_k)$ and X_k are shown by the dotted curves in Figures 8.5 , 8.7 and 8.8 . The solid curves show the actual coordinates for rays traced through the same entrance pupil points. Figures 8.7 (a), (b) and (c) are plots for fans of meridional rays at 10°, 15° and 20° respectively. Figures 8.8 (a) and 8.8 (b) show plots for skew fans with $Y_1 = 0$. Figure 8.9 is a plot of the fractional distortion of the lens as a function of the object field angle. The results show that the actual distortion is slightly more positive than predicted from the third order theory.

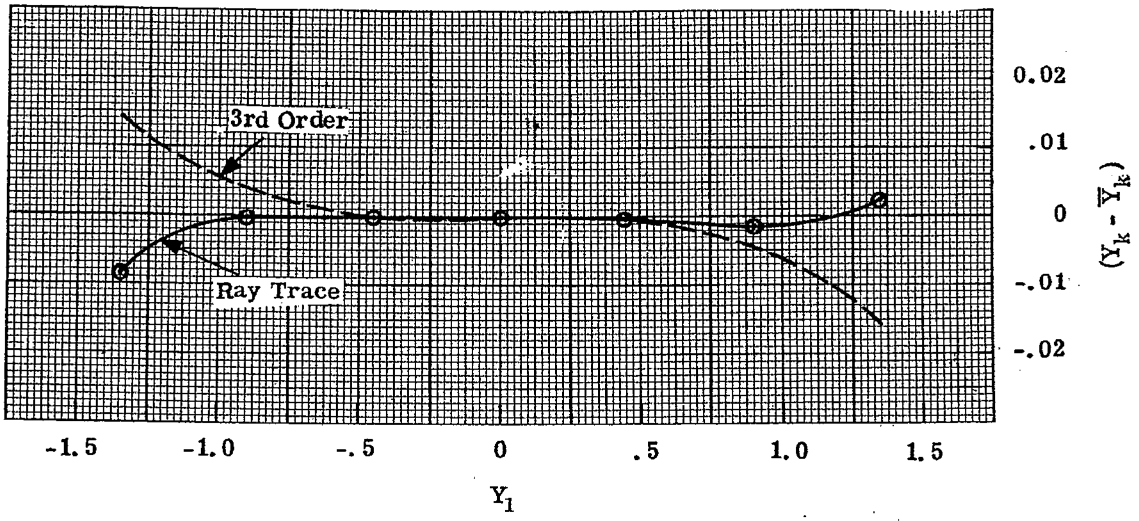
8.6.2.2 Finally in Figure 8.10 the slopes of all the curves at the chief ray are indicated. (Slopes are proportional to t_k). Curves of this type are called field curves. The points on the curves show the longitudinal distances from the paraxial image plane to the focus of rays close to the chief ray. The three third order field curves were found from surface contributions. The remaining two, the tangential field curve and the sagittal field curve, were obtained by graphically determining the slope of the meridional and skew ray plots respectively. These are shown in Figure 8.7, and in Figure 8.8. The third order tangential and sagittal field curves may be calculated by differentiating Equations (6) and (7) with respect to Y_1 and X_1



(c) 20° Obliquity



(b) 15° Obliquity



(a) 10° Obliquity

Figure 8.7 - Meridional ray plots.

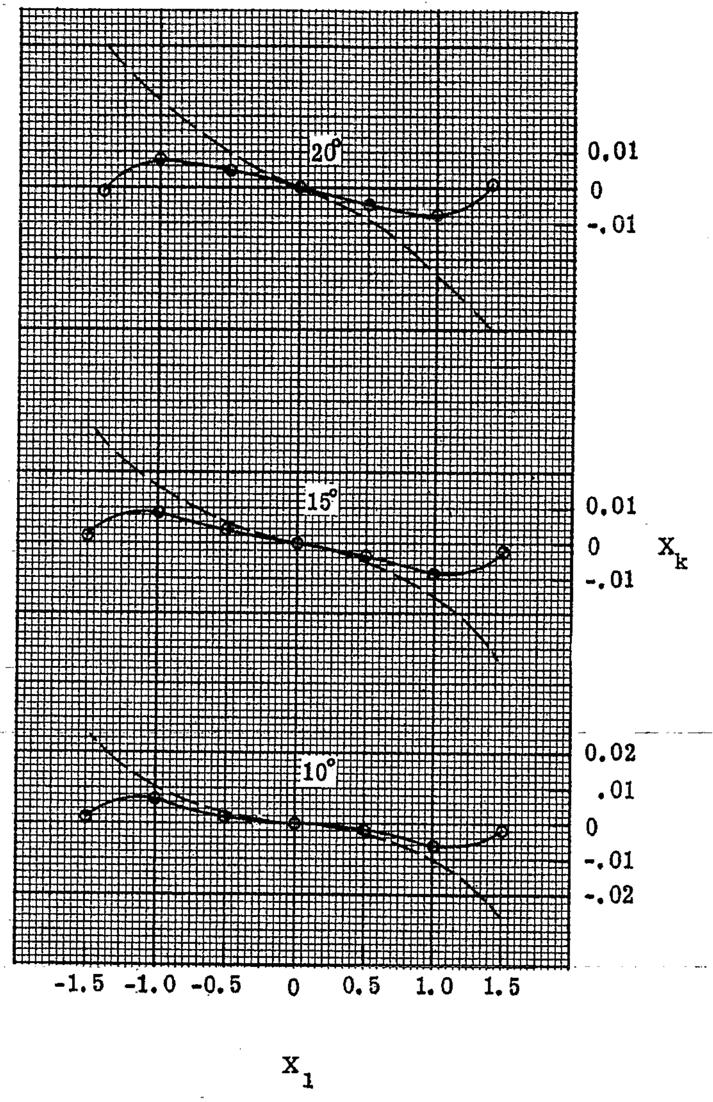


Figure 8.8 (a) - Skew ray image coordinates (X_k).

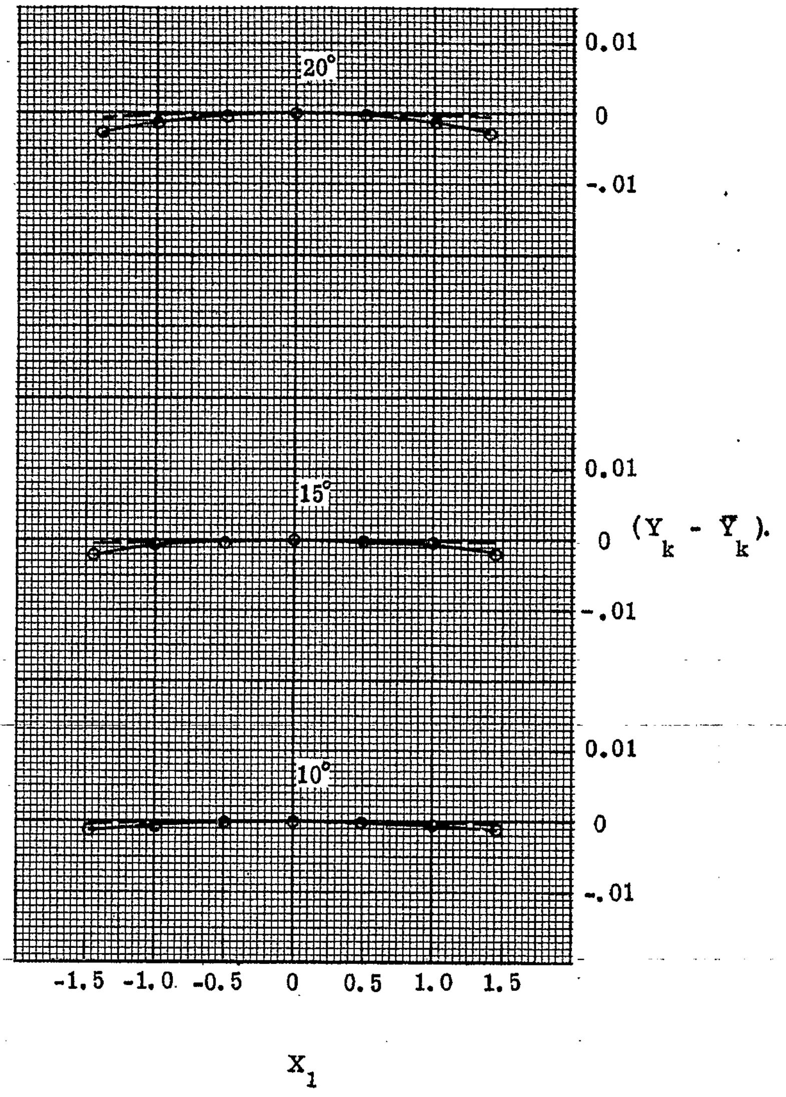


Figure 8.8 (b) - Skew ray image coordinates $(Y_k - \bar{Y}_k)$.

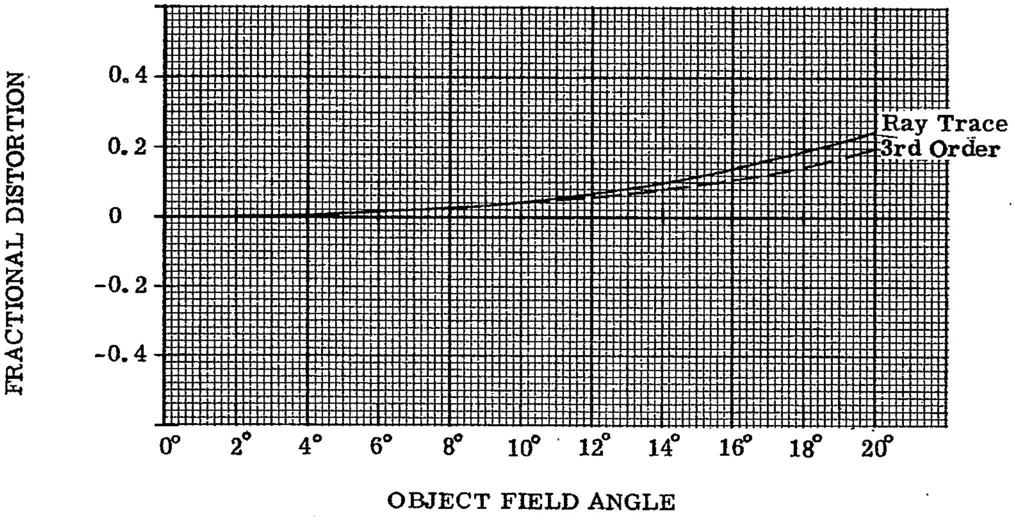


Figure 8.9 - Fractional distortion as a function of field angle.

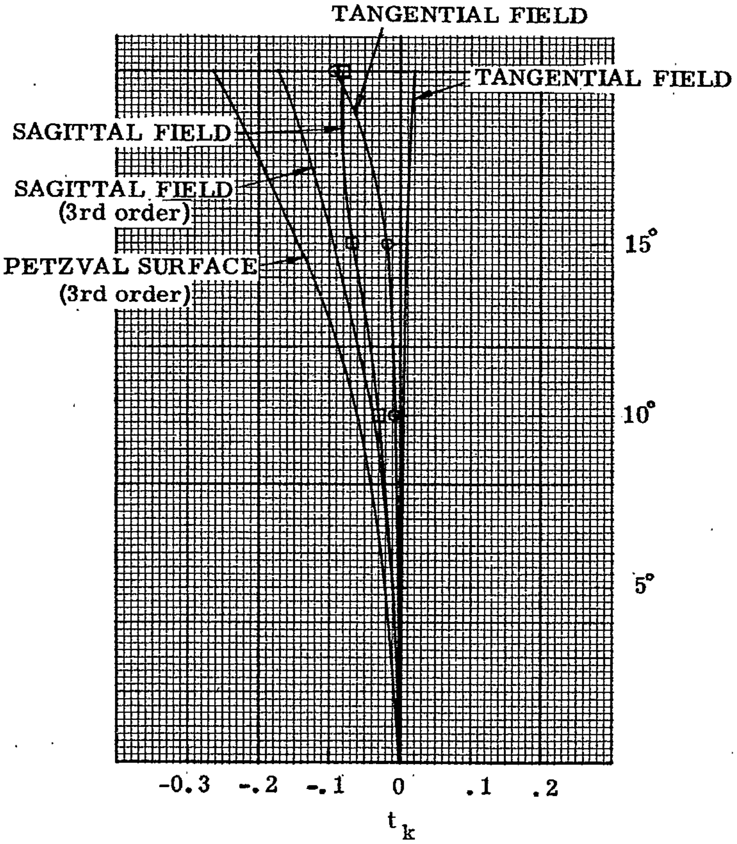


Figure 8.10 - Field curves for a triplet.

respectively and evaluating at $Y_1 = X_1 = 0$. The B and F terms drop out, leaving only the third terms. The final equation for the tangential fan is

$$t_{kT} = \frac{1}{2(n_{k-1}\,u_{k-1}^2)} \left[\Sigma (3C + P\Phi^2) \left(\frac{L_o}{M_o\,\bar{u}_o} \right)^2 \right] . \qquad (9)$$

For the sagittal fan, the equation is

$$t_{kS} = \frac{1}{2(n_{k-1}\,u_{k-1}^2)} \left[\Sigma (C + P\Phi^2) \left(\frac{L_o}{M_o\,\bar{u}_o} \right)^2 \right] . \qquad (10)$$

The Petzval surface curve is determined by the equation

$$t_{kP} = \frac{1}{2}(3t_{kS} - t_{kT}) = \frac{1}{2(n_{k-1}\,u_{k-1}^2)} \left[\Sigma P \Phi^2 \left(\frac{L_o}{M_o\,\bar{u}_o} \right)^2 \right] .$$

Comparing this with Equation (9) and (10), it is clear that for $C = 0$, $t_{kP} = t_{kT} = t_{kS}$.

8.7 CALCULATION OF THE THIRD ORDER CONTRIBUTIONS B, F, C, E AND P

8.7.1 Basic formulae. The method for calculating B, surface by surface, was explained in Paragraph 8.5.2.2, and a sample calculation was given in Table 8.2. The coefficients F, C and E are calculated, surface by surface, by using the data of both the axial and chief paraxial rays. The formulae are:

$$B = S i^2 \qquad (4)$$

$$F = S i \bar{i} \qquad (11)$$

$$C = S \bar{i}^2 \qquad (12)$$

$$E = \bar{S} i \bar{i} + \Phi(\bar{u}_{-1}^2 - \bar{u}^2)^* . \qquad (13)$$

P is calculated for each surface from the equation

$$P = \frac{c(n_{-1} - n)}{n_{-1}\,n} . \qquad (14)$$

As in the case of Equations (3), (4) and (5), Equations (6), (7), (8), (11), (12), (13) and (14) are derived from the Seidel expressions for coma, astigmatism, distortion and Petzval curvature.

8.7.2 Calculation of aberrations. These surface contributions have been worked out, surface by surface, in the sample problem shown in Table 8.2. The individual surface contributions, when summed up for all the surfaces, may be inserted in Equations (6) and (7) to evaluate the third order polynomials.

8.7.3 Fourth order aspheric effects. A fourth order aspheric deformation term on a surface introduces the following amounts of third order aberrations,

$$B = 8(n_{-1} - n)\,e\,y^4 \qquad (4a)$$

$$F = B\bar{y}/y \qquad (11a)$$

$$C = B(\bar{y}/y)^2 \qquad (12a)$$

$$E = B(\bar{y}/y)^3 . \qquad (13a)$$

Note that the aspheric deformation term introduces aberrations independent of the curvature of the surface. It introduces no first order chromatic effects or Petzval contribution.

---

* $\bar{S}$ is calculated from Equation (5) using data from the chief ray.

8.7.4 The value of using third order aberration coefficients.

8.7.4.1 Inspection of the ray tracing data in Figures 8.5, 8.7 and 8.8 shows that the third order aberration polynomial does not predict the true aberration accurately for large apertures or field angles. However, third order aberration theory is extremely valuable. Even with present day computers, it is almost essential for a designer to calculate the third order aberrations of a system under consideration. Third order aberration theory provides target values for the designer; the third order aberrations must be within fairly narrow regions in order to obtain a satisfactory design. It is then up to the designer to find a layout which will lie within this third order region, but which will either balance or reduce the higher order aberrations.

8.7.4.2 Third order surface contributions provide the designer with a means for understanding a lens. He knows that the aberrations should be corrected with evenly balanced third order contributions. In other words, the third order contributions of a single aberration should be approximately equal numerically, but have alternating signs so that the sum is small. A large third order aberration on a surface will introduce a large higher order aberration of the same sign. Hence, the third order aberrations should be kept small. It is surprising how well one can control higher order aberrations through the use of third order calculations by remembering the following recommendations.

(1) Try to find the required third order solution with small, evenly distributed aberration contributions. It is seldom advisable to introduce a large contribution on one surface to cancel out several small ones due to other surfaces.

(2) Try to avoid large angles of incidence. The angle of incidence strongly affects the magnitude of higher order aberration contributions.

(3) If a given surface introduces a large amount of any third order aberration, try to correct this by another surface as nearby as possible. The reason for this is that a surface introducing large amounts of third order aberrations also introduces a series of higher order aberrations. If the third order aberrations are corrected by a neighboring surface, the higher order aberrations tend to cancel one another, but if correction is done at some other part of the optical system, the higher order aberrations will not necessarily cancel. For example, if a large amount of spherical aberration is introduced at a position in the system where $\bar{y}/y = k$, then this aberration should be corrected at a surface as close as possible to the position where $\bar{y}/y = k$. It may often be impossible, in a given design, to make the ideal correction, but it is an important step in design procedure to make the attempt. One of the main reasons that aspheric surfaces are so valuable, is that they do allow the introduction of aberration at nearly any place in the optical system, without upsetting the distribution of focal lengths of the different elements needed to correct for color and Petzval field curvature.

8.8 AFOCAL OPTICAL SYSTEMS

8.8.1 Third order polynomial. In telescopic systems, where both the object and image are at infinity, it is convenient to plot the tangents of the angles which the emerging rays make with the optical axis, versus the coordinates (X_1, Y_1) of the entering rays. The meridional ray $(X_1 = 0$ and Y_1 arbitrary$)$ data are plotted as $\left(\frac{L_{k-1}}{M_{k-1}} - \frac{\bar{L}_{k-1}}{\bar{M}_{k-1}}\right)$ versus Y_1. The skew ray $(Y_1 = 0$ and X_1 arbitrary$)$ data are plotted as two curves:

$$\frac{K_{k-1}}{M_{k-1}} \quad \text{versus} \quad X_1 ,$$

and

$$\left(\frac{L_{k-1}}{M_{k-1}} - \frac{\bar{L}_{k-1}}{\bar{M}_{k-1}}\right) \quad \text{versus} \quad X_1 .$$

The third order polynomial may then be written as in Equations (6) and (7) by making the following substitutions:

$$Y_k = -\frac{y_{k-1}}{u_{k-1}} \tan U_{k-1} \qquad X_k = -\frac{y_{k-1}}{u_{k-1}} \frac{K_{k-1}}{M_{k-1}}$$

and

$$\bar{Y}_k = -\frac{y_{k-1}}{u_{k-1}} \tan \bar{U}_{k-1} .$$

Equations (6) and (7) then become

$$\left(\frac{L_{k-1}}{M_{k-1}} - \frac{\bar{L}_{k-1}}{\bar{M}_{k-1}}\right) = \frac{1}{2(n_{k-1}\, y_{k-1})} \left[\Sigma B\,(Y_1^2 + X_1^2)\,\frac{Y_1}{y_1^3} + \Sigma F\,\frac{(3Y_1^2 + X_1^2)}{y_1^2}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right) + \Sigma(3C + P\Phi^2)\,\frac{Y_1}{y_1}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2\right] + O(5)\,, \tag{15}$$

and

$$\frac{K_{k-1}}{M_{k-1}} = \frac{1}{2\,(n_{k-1}\, y_{k-1})} \left[\Sigma B\,(Y_1^2 + X_1^2)\left(\frac{X_1}{y_1^3}\right) + \Sigma F\,\frac{(2\,Y_1\;X_1)}{y_1^2}\left(\frac{\bar{Y}_o}{\bar{y}_o}\right) + \Sigma(C + P\Phi^2)\left(\frac{X_1}{y_1}\right)\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2\right] + O(5)\,. \tag{16}$$

8.8.2 Spot diagram.

8.8.2.1 In an ideal, aberration-free afocal system, the emergent rays from the $k - 1$ surface are parallel. In a real afocal system these rays are almost parallel. The intersection of these rays with a plane would give a series of points more or less evenly spaced; the points would not be concentrated, as in a spot diagram (see Figure 8.1 (b)), and it would be difficult to interpret the diagram in the same way as in the case of the spot diagram.

8.8.2.2 It is possible to concentrate these almost parallel emergent rays and make a spot diagram for an afocal system by adding a hypothetical aberration-free thin lens at the rear of the system with any desired focal length. This is effectively done simply by changing the coordinates for each ray on the last surface of the system to zero. The rays then proceed to the final focal plane of the aberration-free lens from this point at the same angles because they pass through the center (coinciding nodal points) of the thin lens. The distance to the image plane is the arbitrary focal length of this lens. The spread of the points from a single, concentrated spot, is an indication of the non-parallelism of the emergent rays. This in turn is an indication of the aberrations of the afocal system.

8.9 STOP SHIFT EQUATIONS

8.9.1 General. The aim of a lens designer is to minimize the aberrations of the optical system within the specifications of f - number and field of view. It is clear by Equations (4), (5), (11), (12), (13) and (14) that the third order coefficients depend on index, curvature, and thickness. By Equations 6-(34) and 6-(35), the first order chromatic coefficients also depend on these parameters. But the occurrence of $\bar{i}$, $\bar{S}$, and $\bar{u}_{-1}$, in Equations (11), (12), (13) and 6-(35), show that the oblique aberrations (coma, astigmatism, distortion, and lateral color) depend on the position of the aperture stop as well. Hence it is necessary for the designer to know the effect of the stop position on these aberrations.

8.9.2 Aberration polynomial for a shifted chief ray.

8.9.2.1 The aberration polynomials shown in Equations (6) and (7) are calculated from the coefficients B , F , C and E which are determined by tracing an axial and an oblique chief paraxial ray. It is possible to compute the aberration polynomial for any other paraxial chief ray. The term other paraxial chief ray, or shifted chief ray, refers to another ray which crosses the optical axis at the new pupil points. Hence shifting the aperture stop results in a new ray becoming the (shifted) chief ray. Suppose we wish to write down the aberration polynomial for a paraxial chief ray which passes through the original entrance pupil at a height of $\bar{y}^*$. A ray from object point $\bar{Y}_o$ passing through the original entrance pupil at a height of Y_1 will be at a height of Y_1' in the original entrance pupil above the new chief ray. Figure 8.11 shows that $Y_1 = Y_1' + \bar{Y}_1^*$.

8.9.2.2 Equation (6) may be written now in terms of Y_1' . The distortion term in Equation (8) is added to Equation (6) to give the aberration $Y_k - \bar{Y}_o\, m$. For an object of height $\bar{Y}_o$, it can be seen by similar triangles that $\bar{Y}_1^*$ is given by

$$\bar{Y}_1^* = \bar{y}_1^* \left(\frac{\bar{Y}_o}{\bar{y}_o}\right) .$$

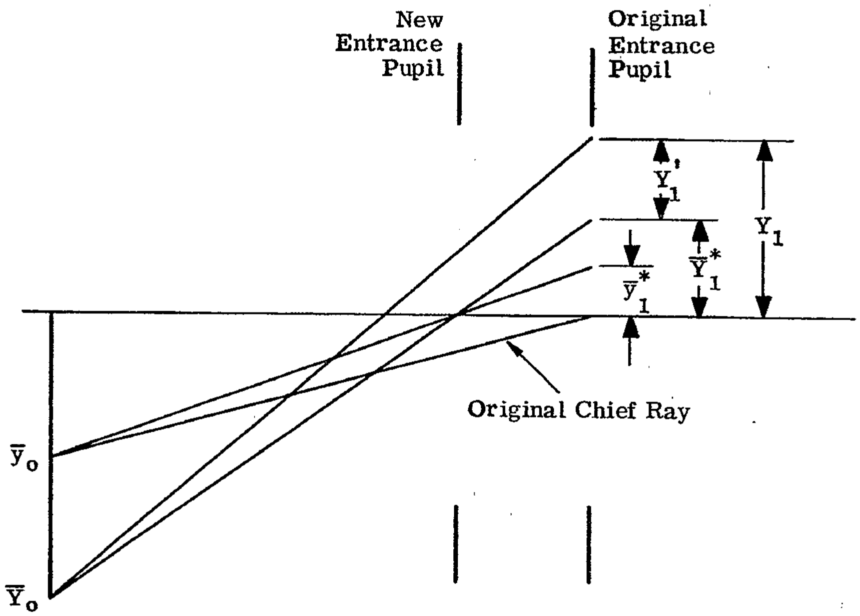


Figure 8.11 - Application of stop shift equations.

Then the height Y_1' of the ray above the new chief ray is

$$Y_1' = Y_1 - \bar{Y}_1^* = Y_1 - \bar{y}_1^* \left(\frac{\bar{Y}_o}{\bar{y}_o}\right) .$$

By substituting this expression into the sum of Equations (6) and (8), and by using the relation,

$$Q = \bar{y}_1^* / y_1 ,$$

it is possible to arrive at the equation

$$(Y_k - \bar{Y}_o m) = -\frac{1}{2(n_{k-1} u_{k-1})} \left[\Sigma B (Y_1'^2 + X_1^2)\left(\frac{Y_1'}{y_1^3}\right) + (Q \Sigma B + \Sigma F)\left(\frac{3Y_1'^2 + X_1^2}{y_1^2}\right)\left(\frac{\bar{Y}_o}{\bar{y}_o}\right) + \right.$$
$$[3Q^2 \Sigma B + 6Q \Sigma F + \Sigma(3C + P\Phi^2)] \left(\frac{Y_1'}{y_1}\right) \left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2 +$$
$$\left. [Q^3 \Sigma B + 3Q^2 \Sigma F + Q\Sigma(3C + P\Phi^2) + \Sigma E]\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^3 \right] . \qquad (17)$$

8.9.3 Third order aberration coefficients. Equation (17) is the aberration polynomial with a shifted chief ray and therefore a shifted entrance pupil. If this equation is compared with the sum of Equations (6) and (8) it has a similar form. In Equation (17) the original aberration coefficients ΣB, ΣF, ΣC, ΣE and ΣP have been replaced by linear combinations of these coefficients. Since the aberration polynomial has the same form it can be said that the third order coefficients have changed to new values. The new third order coefficients will be indicated with a superscript *. By comparing Equation (17) with Equations

(6) and (8) it follows that

$$\Sigma B^* = \Sigma B , \tag{18}$$

$$\Sigma F^* = Q \Sigma B + \Sigma F , \tag{19}$$

$$\Sigma C^* = Q^2 \Sigma B + 2Q \Sigma F + \Sigma C , \tag{20}$$

$$\Sigma E^* = Q^3 \Sigma B + 3Q^2 \Sigma F + Q \Sigma (3C + P \Phi^2) + \Sigma E , \tag{21}$$

and

$$\Sigma P^* = \Sigma P . \tag{21a}$$

8.9.4 First order aberration coefficients. Using Equation 6-(39), to complete the list of changes of the aberration coefficients as the chief ray is changed, the chromatic coefficients then become

$$a^* = a , \tag{22}$$

and

$$b^* = Q a + b . \tag{23}$$

8.9.5 Use of the stop shift equations.

8.9.5.1 These equations, (18-23), are often called the stop shift equations. They are extremely useful and will be referred to many times in later sections. They enable the designer to predict how the third order coefficients change with the choice of the chief ray. Again we see that any two paraxial rays traced through the lens are sufficient for all third order calculations on the system. If the oblique chief ray traced through the system turns out not to pass through the center of the new aperture stop, it is possible to use the stop shift formulae to compute the third order coefficients for the chief ray that does pass through.

8.9.5.2 The designer should note that Equation (17) uses the aperture variable Y_1'. As shown in Figure 8.11, Y_1' is the height of a general ray above the new chief ray in the original entrance pupil. The height of this general ray above the chief ray in the new entrance pupil will not be Y_1' if the object is at a finite distance. In order to write the polynomial in terms of Y_1 one must account for the magnification between the original and the new entrance pupil. Now it should be noted that the polynomial involves the ratio of Y_1' / y_1. It turns out that the corresponding ratio for the new entrance pupil has the same value. Therefore the aberration polynomial, Equation (17), can be used with Y_1' and y_1 as coordinates in the new entrance pupil.

8.10 THIN LENS ABERRATION THEORY

8.10.1 Third order coefficients. It is possible to combine the two surface contributions of a thin lens in air and obtain simple expressions for the third order aberrations. By assuming that the lens is thin, the values of y for the axial paraxial ray are the same on the two surfaces. If it is further assumed that the lens is the aperture stop (and hence the entrance and exit pupils), the oblique paraxial chief ray passes through the center of the thin lens at $\bar{y} = 0$. The third order aberration coefficients of the thin lens are then

$$B = \alpha_1 + \alpha_2 c_1 + \alpha_3 c_1^2 , \tag{24}$$

$$F = \beta_1 + \beta_2 c_1 , \tag{25}$$

$$C = - \phi \Phi^2 , \tag{26}$$

$$E = 0 , \tag{27}$$

and

$$P = - \frac{\phi}{n} . \tag{28}$$

The constants of the new equations are:

$$\alpha_1 = -\frac{\phi y^4}{n}\left[(3n+2)\left(\frac{u_{-1}}{y}\right)^2 + \left(\frac{\phi n}{n-1}\right)^2 n - \frac{\phi n}{n-1}(3n+1)\frac{u_{-1}}{y}\right], \quad (29)$$

$$\alpha_2 = -\frac{\phi y^4}{n}\left[(4n+4)\left(\frac{u_{-1}}{y}\right) - \left(\frac{\phi n}{n-1}\right)(2n+1)\right], \quad (30)$$

$$\alpha_3 = -\frac{\phi y^4}{n}(n+2), \quad (31)$$

$$\beta_1 = \frac{\phi \Phi y^2}{n}\left[(2n+1)\frac{u_{-1}}{y} - \left(\frac{\phi n}{n-1}\right) n\right], \quad (32)$$

$$\beta_2 = \frac{\phi \Phi y^2}{n}(n+1), \quad (33)$$

$$\phi = \frac{(u_{-1} - u)}{y} = \frac{1}{f'} \text{ (from Equation 6-(24)) }, \quad (34)$$

and

$$\Phi = \bar{y}\, u_{-1} - y\, \bar{u}_{-1} = \bar{y}\, u - y\, \bar{u}, \quad (35)$$

where u_{-1} is the angle of the axial paraxial ray in air on the left hand side of the thin lens and u is the angle of this ray in air on the right hand side, n is the index of the lens,and c_1 is the curvature of the first surface.

8.10.2 Limitations; comparison with thick lens results.

8.10.2.1 Equations (24) through (28) are valid for any thin lens in air at any position in a system provided $\bar{y} = 0$ at the lens. If the value of $\bar{y}$ is not zero it is necessary to calculate B^*, F^*, C^*, E^*, and P^* from the stop shift equations (18) through (21a). With the proper substitution it can be shown that,

$$B^* = \alpha_1^* + \alpha_2^* c_1 + \alpha_3^* c_1^2, \quad (36)$$

$$F^* = \beta_1^* + \beta_2^* c_1 + \beta_3^* c_1^2, \quad (37)$$

$$C^* = \gamma_1^* + \gamma_2^* c_1 + \gamma_3^* c_1^2, \quad (38)$$

$$E^* = \delta_1^* + \delta_2^* c_1 + \delta_3^* c_1^2. \quad (39)$$

The coefficients of these quadratic equations are as follows:

$$\alpha_1^* = \alpha_1, \quad \alpha_2^* = \alpha_2, \quad \alpha_3^* = \alpha_3, \quad (40)$$

$$\beta_1^* = Q\alpha_1 + \beta_1, \quad (41)$$

$$\beta_2^* = Q\alpha_2 + \beta_2, \quad (42)$$

$$\beta_3^* = Q\alpha_3, \quad (43)$$

$$\gamma_1^* = Q^2\alpha_1 + 2Q\beta_1 - \phi\Phi^2, \quad (44)$$

$$\gamma_2^* = Q^2\alpha_2 + 2Q\beta_2, \quad (45)$$

$$\gamma_3^* = Q^2\alpha_3, \quad (46)$$

$$\delta_1^* = Q^3\alpha_1 + 3Q^2\beta_1 - Q(3n+1)\frac{\phi\Phi^2}{n}, \quad (47)$$

$$\delta_2^* = Q^3\alpha_2 + 3Q^2\beta_2, \quad (48)$$

and

$$\delta_3^* = Q^3\alpha_3. \quad (49)$$

8.10.2.2 To illustrate the use of these equations a sample calculation for all the thin lens coefficients is presented in Table 8.4. In this example, the calculations were made on the thin lens illustrated in Table 6.13. Table 8.2 shows the same lens system with thickness added. The thin lens equations were used with

$c_1 = 0.253$ for lens (a),

$c_1 = -0.200$ for lens (b),

and

$c_1 = 0.050$ for lens (c).

8.10.2.3 Table 8.3 lists a comparison between the third order aberration coefficients calculated from the thin lens equations and those calculated from the surface contributions of the thick lens. The differences between the coefficients is due to the thicknesses introduced in the sample shown in Table 8.2. Slight differences are also due to the differences in c_1.

8.10.2.4 In the following sections it will be demonstrated how the thin lens equations are used in the preliminary design of a lens system.

Lens and Coefficient		Thin Lens Formula	Thick Lens Formula
Lens (a) $C_1 = 0.253$	B	-0.0365	-0.0355
	F	0.0241	0.0221
	C	-0.0647	-0.0604
	E	0.1046	0.0962
	P	-0.1021	-0.1024
Lens (b) $C_1 = -0.200$	B	0.0722	0.0731
	F	-0.0340	-0.0352
	C	0.0895	0.0919
	E	-0.0182	-0.0195
	P	0.1770	0.1759
Lens (c) $C_1 = 0.050$	B	-0.0443	-0.0440
	F	0.0136	0.1029
	C	-0.0290	-0.0272
	E	-0.0776	-0.0746
	P	-0.1124	-0.1135

Table 8.3 - Comparison between third order aberrations calculated from thin lens equations and from individual surface contributions of a thick lens.

Quantity	Lens (a)	Lens (b)	Lens (c)
y^2	2.25000	1.28989	1.56630
y^4	5.06250	1.66380	2.45329
$-\phi y^4/n$	-0.51678	0.29456	-0.27574
$\phi\Phi$	-0.09029	0.15669	-0.09942
$\phi\Phi y^2/n$	-0.12541	0.12468	-0.09612
$-\phi\Phi^2$	-0.04930	0.08555	-0.05428
$n+1$	2.6200	2.6210	2.6200
$n+2$	3.6200	3.6210	3.6200
$2n+1$	4.2400	4.2420	4.2400
$3n+1$	5.8600	5.8630	5.8600
$3n+2$	6.8600	6.8630	6.8600
$4n+4$	10.4800	10.4840	10.4800
u_{-1}/y	0	-0.21841	0.06223
$\phi n/n-1$	0.43210	-0.74911	0.47576
$(3n+2)\left(\frac{u_{-1}}{y}\right)^2$	0	0.32739	0.02656
$\left(\frac{\phi n}{n-1}\right)^2 n$	0.30246	0.90964	0.36668
$-\frac{\phi n}{n-1}(3n+1)\frac{u_{-1}}{y}$	0	-0.95926	-0.17348
α_1	-0.15631	0.08182	-0.06060
$(4n+4)\frac{u_{-1}}{y}$	0	-2.28981	0.65213
$-\frac{\phi n}{n-1}(2n+1)$	-1.83209	3.17771	-2.01721
α_2	0.94679	0.26153	0.37641
α_3	-1.87075	1.06659	-0.99817
$(2n+1)u_{-1}/y$	0	-0.92650	0.26384
$-\left(\frac{\phi n}{n-1}\right)n$	-0.70000	1.21430	-0.77073
β_1	0.08778	0.03588	0.04872
β_2	-0.32856	0.32680	-0.25183
Q	-0.53333	-0.06268	0.50844
$\beta_1{}^*$	0.17115	0.03076	0.01791
$\beta_2{}^*$	-0.83352	0.31040	-0.06045
$\beta_3{}^*$	0.99773	-0.06686	-0.50752
$\gamma_1{}^*$	-0.18740	0.08138	-0.02040
$\gamma_2{}^*$	0.61978	-0.03994	-0.15878
$\gamma_3{}^*$	-0.53212	0.00419	-0.25804
$\delta_1{}^*$	0.19373	-0.01899	-0.07001
$\delta_2{}^*$	-0.42400	-0.00379	-0.14584
$\delta_3{}^*$	0.28380	-0.00026	-0.13120

Table 8.4 - Calculation of the thin lens coefficients for the thin lens shown in Table 6.13.

9 METHOD OF LENS DESIGN

9.1 · THE PROCESS OF DESIGNING A LENS SYSTEM

9.1.1 Introduction. The formulae used to design a lens system have now been presented. Ray trace equations were derived in Section 5. Their use in first order design and in aberration analysis were discussed in Sections 6 and 8. In the present section a systematic method for the design of lens systems will be described, and this method will be illustrated with the design of a triplet flat field lens in Section 10.

9.1.2 Approach. The design of a lens system at the present state of the art is an iterative procedure. Certain steps in the procedure are repeated until a satisfactory design is attained. In this sense, lens design involves a trial and error procedure. At present (1962), direct methods of design, proceeding from the desired specifications to the specific lens, do not exist. The following steps are the basic elements of the iterative procedure.

(1) Select a lens type.

(2) Find a first order thin lens solution.

(3) Find a third order thin lens solution.

(4) Find a thick lens solution, and calculate first order and third order aberrations.

(5) Trace a few selected meridional and skew fans.

(6) Adjust third order coefficients to balance higher order aberrations, and repeat steps 5 and 6 until the balance between third and higher order aberrations agrees with desired specifications, or at least is reasonable.

(7) Trace additional fans of skew rays; make spot diagrams and calculate the energy distribution.

(8) Evaluate the image.

(9) Return to a previous step and repeat the process until evaluation indicates desired performance. Which step to return to depends on the problem. The most usual procedure is to return to step (4), but often the designer must return to step (1).

9.2 DESCRIPTION AND ANALYSIS OF THE BASIC PROCEDURE

9.2.1 Step 1 - Selection of a lens type.

9.2.1.1 In order to select the type of lens to be designed, the designer must first survey the complete lens problem. He attempts to equate it to one of the simple basic optical systems. He asks if this is a magnifier problem, a microscope, a telescope or a camera lens. After deciding upon the basic system, he then proceeds to make a layout using simple theory as illustrated in Section 7. This analysis thus generates a possible arrangement picture of how the axial and oblique rays will pass through the system.

9.2.1.2 Suppose, for example, that the system to be designed is a telescope. Given the magnifying power, field of view and over-all instrument length, a designer may conclude that the telescope should consist of an objective, a prism erecting system and an eyepiece. From the preliminary analysis he concludes that the objective must work at $f/3.5$ and the eyepiece must cover a half field of 30°. Looking over objective designs (for example, see Section 11) he may then compute the field curvature for the system and conclude that he will use an objective like the one illustrated in Figure 11.7, and, since the eyepiece must cover a half field of 30°, an Erfle type appears to be a logical choice. Inspection of the eyepieces shown in Section 14 discloses that the Erfle is the simplest design. It represents a good starting point.

9.2.1.3 Other factors may influence the designer's choice. Compatability with other systems, existing hardware, economics or delivery schedule are all valid considerations. Thus, unfortunately for the beginner, this step in the procedure is difficult and requires the most experience. As the process proceeds, the steps become more automatic and less dependent on experience. This means that the beginner finds it difficult to get started and it means that the designer instructing must say in effect at the beginning, "Let us start with a lens of such and such a type. Later I may be able to show you why I picked this particular type of solution." This approach to a problem does not appeal to the analytic mind but at present there is no other way to approach the problem. It would be nice if one could work from the specifications of the image, back to the design required, but there are only very limited procedures which will enable one to establish what lens type is needed for a particular problem. In Sections 10, 11, 12, 13, and 14 the performance and limitations of several types of lenses will be described which it is hoped will help a beginner select the type of lens.

9.2.1.4 The prime accomplishment of this step is the designer's decision to choose a certain lens type to perform a specified function in the system. Thus a starting point is established from which computation and evaluation can proceed. This step, baffling as it is to the beginner, is really the most creative part of the design, and, as experience is gained, this is the part of the design that intrigues the designer and gives him a chance to exercise judgement, which is what humans usually enjoy.

9.2.2 Step 2 - The first order thin lens solution. Once the lens type has been decided on, the next step is to solve the algebraic equations to determine the individual focal lengths and spacings of the elements. It greatly simplifies the procedure to assume that the lenses are thin. At this stage of the problem, there are usually conditions that must be satisfied in the passage of the axial paraxial ray, and the oblique paraxial ray. The entrance and exit pupils may have to be located at special positions, and their sizes may be given. The focal length and back focal length may be specified. It is also necessary to adjust the axial and lateral color, and Petzval sum to appropriate values. The passage of the oblique chief ray has an effect on the distortion. For simple systems it often is possible to write down algebraic equations relating the parameters of the system (ϕ, t, n) and the required conditions to be satisfied, but very often the algebra becomes so complex that graphical or linear approximations are required to find the solutions. The problem basically amounts to trying to solve a set of non-linear equations. Sometimes there are more equations than variables, in other instances the reverse may be true. One can spend a great deal of time on the algebra at this stage of the design. Often, the most sensible procedure is to resort to a systematic trial and error solution. This method will be illustrated in Section 10.2

9.2.3 Step 3 - The third order thin lens solution. By making the thin lens aberration coefficient calculations illustrated in Table 8.4, it is possible to obtain sets of second degree algebraic equations relating the first curvatures of the lenses and the aberrations. Again, in simple systems these can sometimes be solved algebraically or graphically. As a matter of fact, if these equations cannot be solved algebraically there is little justification for using the thin lens approximations, for one can as readily apply the trial and error methods to thick lenses using the surface contribution calculations shown in Table 8.2. By properly choosing the position of the aperture stop it is possible to greatly simplify the equations. The following reasoning is used in the preliminary design. In the preliminary third order design the aberrations are usually all made equal to zero. Equations 8-(18) through 8-(21) show us, that if B, F, C, E and P are all set to zero, then B*, F*, C*, and E* will all be zero. This tells us that the location of the stop position has no effect on the aberrations. Then it is advisable to choose the chief ray to pass through the center of one of the lenses. By so doing, the aberrations for this lens are given by Equations 8-(24) through 8-(28). This eliminates the calculation of E, the C is constant, and F varies linearly with c_1. In practice, it helps to use this procedure even if small residual aberrations are to be left in the system.

9.2.4 Step 4 - First and third order aberrations of a thick lens.

9.2.4.1 During this step in the design, calculations of the type shown in Table 8.2 are made to determine the first and third order aberrations of the lens with actual thicknesses. If the thin lens theory has been worked out completely, then values for the curvatures and the desired angles of the axial and oblique rays are known. Now, the procedure of introducing thicknesses changes all the first order and third order aberrations. The next problem is to modify the thick lens solution to achieve the desired aberrations.

9.2.4.2 Some designers have procedures for computing the positions of the principal planes of each individual element. Then the thick lens system is set up so that the first curvatures of each of the lenses are the same as for each of the thin lenses, and the angles the axial ray makes with the axis is the same as for the thin lenses. Finally, the spaces between the lenses are adjusted to make the spacings between the image principal plane (P_{2a}) and the next object principal plane (P_{1b}) of the thick lenses equal to the spacing between the thin lenses.

9.2.4.3 The designer should not spend too much time trying to adjust the spacings in this way since there is no direct and easy way to set up a thick lens equivalent of the thin lens. The procedure just described always fails to keep all the aberrations the same as for the thin lens; some changes in the power distribution are necessary.

9.2.4.4 If the designer is setting up for the first time a thick lens from thin lens data, there is really very little point in trying to make the thick lens aberrations exactly equal to the thin lens aberrations. The reason for this is that until one has ray traced a design, and determined the magnitude of the higher order aberrations, it is not possible to tell just what third order aberrations are needed to balance out those of a higher order. Usually, a perfectly satisfactory way to set up a thick lens from thin lens data is to assume the positions of the principal planes, from a simple sketch of the lens, using curvatures from the thin lens solution and thicknesses from 10.4.

9.2.4.5 A major problem in lens design is the problem of adjusting a thick lens to arrive at some definite third order aberrations. This can be done by a trial and error method if some information is known about how the aberrations vary with parameter changes. Sometimes the information in the form of curves for the thin lenses provides indications to the designer which help him decide how to adjust the thick lens to find a solution.

9.2.4.6 The problem of adjusting a thick lens system resolves itself into the problem of solving a set of simultaneous equations. One can systematically change one parameter at a time and recalculate all the total aberrations of the new system. By finding the differences in the total aberrations due to the parameter change, it is possible to compute the parameter differential for all the third and first order aberrations. This method will now be discussed in detail.

9.2.4.7 Since B, F, C, E, P, a, and b are functions of all the system parameters, it is possible to write

$$\Delta\Sigma B = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma B}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma B}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma B}{\partial n}\right)_j \Delta n_j \right], \qquad (1)$$

$$\Delta\Sigma F = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma F}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma F}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma F}{\partial n}\right)_j \Delta n_j \right], \qquad (2)$$

$$\Delta\Sigma C = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma C}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma C}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma C}{\partial n}\right)_j \Delta n_j \right], \qquad (3)$$

$$\Delta\Sigma E = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma E}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma E}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma E}{\partial n}\right)_j \Delta n_j \right], \qquad (4)$$

$$\Delta\Sigma P = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma P}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma P}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma P}{\partial n}\right)_j \Delta n_j \right], \qquad (5)$$

$$\Delta\Sigma a = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma a}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma a}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma a}{\partial n}\right)_j \Delta n_j + \left(\frac{\partial\Sigma a}{\partial \nu}\right)_j \Delta \nu_j \right], \qquad (6)$$

$$\Delta\Sigma b = \sum_{j=1}^{j=k-1} \left[\left(\frac{\partial\Sigma b}{\partial c}\right)_j \Delta c_j + \left(\frac{\partial\Sigma b}{\partial t}\right)_j \Delta t_j + \left(\frac{\partial\Sigma b}{\partial n}\right)_j \Delta n_j + \left(\frac{\partial\Sigma b}{\partial \nu}\right)_j \Delta \nu_j \right]. \qquad (7)$$

9.2.4.8 In order to correct a finite thickness lens system to any desired third and first order aberrations, a designer must, in effect, solve this set of simultaneous equations. Now since B, F, C, E, P, a, and b do not change linearly with parameter changes, these equations will not, in general, provide the correct changes, so the process must be repeated for a series of iterations. Without a large computer it was a hopelessly long procedure to systematically correct a system to a given set of third order values. Therefore, designers had to resort to other techniques. They did this by separating the problem into two parts. First, a solution was found which corrected a, b, and P, with some consideration given to E. Second, this solution was corrected for B, F, and C.

9.2.4.9 The first step, correction of a, b, P, and E, was done by adjusting the focal lengths of the lenses and the spacings between the lenses. Sometimes different coefficients were found for the changes of a, b, and P, and simultaneous equations solved, but the index and dispersion of glass were usually not included because glasses are manufactured in finite steps. Usually designers resorted to a simple trial and error method of adjusting focal lengths, spaces, and glasses. It is surprising how rapidly an experienced designer can adjust variables and arrive at a solution without actually solving the above equations.

9.2.4.10 The second step, correction of B, F, and C, was done by the technique called bending. Lens bending means changing the shape of a lens without affecting its focal length. Equation 6-(22)

gives the expression for the power ϕ of a lens as $(c_1 - c_2)(n-1)$. As long as $c_1 - c_2$ remains constant, c_1 may take on any value without affecting ϕ. If the lens is thin, then bending does not change the angles of the axial and oblique paraxial rays after passage through the lens. If the lens is thick, keeping $(c_1 - c_2)$ constant is not quite the same thing as keeping the focal length fixed because f' depends on t as well as $(c_1 - c_2)$. Usually in bending thick lenses, it is advisable to solve for the second curvature so that the axial ray remains at a constant angle with the optical axis. Bending of a lens has no effect on a, b and P for a thin lens and a very small effect in a thick lens. The bending affects primarily B, F, and C.

9.2.4.11 Therefore, before the widespread use of computers, designers found solutions for given values of B, F, and C by setting up three simultaneous equations. Usually many of the possible degrees of freedom were not used. Experienced lens designers seldom actually solved the equations, but they would keep adjusting the lens by a trial and error method. In the lens designers' slang, the method for finding a solution is jiggle in or poke at it. It is amazing how successfully an experienced designer could jiggle in a design. This method appears to be an art. With experience a designer apparently develops a procedure analogous to solving these equations in his head, by developing a feel for the system.

9.2.4.12 With the modern computer it is now feasible to find automatically a solution of Equations (1) through (7). In Section 10 several examples will be shown illustrating how this is done. Up to the present, the equations solved automatically by the computer have not included the terms with the glass type as a variable. Many problems have been solved using curvatures and thicknesses as variables. The automatic program does essentially the following:

(1) All the first and third order calculations are computed for an initial system. Call this system No. 1.

(2) Each system parameter (c or t) is varied one at a time, and all the first and third order aberrations are calculated for each altered system. The designer may specify which curvatures and thicknesses to change. Each parameter is changed by 0.01% of its initial value.

(3) Differential coefficients are then computed for each variable and aberration. For example,

$$(c_{new} - c_{old})_j = \Delta c_j ,$$

and

$$(\Sigma B_{new} - \Sigma B_{old})_j = \Delta \Sigma B_j .$$

Then

$$\left(\frac{\partial \Sigma B}{\partial c}\right)_j \approx \frac{\Delta \Sigma B_j}{\Delta c_j}$$

(4) When all the differential coefficients are known, the data for the seven equations (1) through (7) are known. The numbers on the left hand side of the equation are found by taking the difference between the aberrations in system No. 1 and the final target (desired) values for the aberrations. For example,

$$\Delta \Sigma B = (\Sigma B_{target} - \Sigma B_1) \text{ etc.}$$

(5) If there are seven variables in the optical system then there will be seven equations with seven unknowns. If there are more variables than equations then the set of equations cannot be uniquely solved. One technique is to impose the condition, that the sum of the squares of the changes in the parameters shall be a minimum. If there are fewer variables than equations then it is not possible to obtain an exact solution. In this case it is customary to solve for a least squares solution. This means a solution is found when the sum of the squares of the differences between the final aberrations and their target values is a minimum.

(6) If the aberrations changed linearly with parameter changes, the target values for the aberrations would be found in one step. However the changes are not usually linear, so the process has to be repeated several times. If the target values for the aberrations are far removed from the initial values, there is the real possibility that this inherently simple procedure will not converge to a solution. Knowledge of the regions of solution is an invaluable aid in helping to select the initial values for system No. 1.

9.2.5 Step 5 - Tracing a few selected meridional and skew fans.

9.2.5.1 After the third order solution is found, the next step is to trace a few selected rays to evaluate the effects of higher order aberrations. The number of rays to trace depends on the stage of the design. On the first ray trace of a new system, only a small number of rays need be traced, but as the design proceeds, additional rays may be necessary for added refinement.

9.2.5.2 One suggested plan for the ray tracing of a design is as follows:

(1) The 0° image. In D light trace three rays at $Y_1 = (Y_1)_{max.}$,

$Y_1 = 0.7\ (Y_1)_{max.}$, $Y_1 = 0.5\ (Y_1)_{max.}$,

where $(Y_1)_{max.}$ is the radius of the entrance pupil. Trace the same rays in F and C light.

(2) If the object is at infinity, trace three meridional fans of rays at angles corresponding to $L_o = (L_o)_{max.}$, $L_o = 0.7\ (L_o)_{max.}$ and $L_o = 0.5\ (L_o)_{max.}$. If the object is at a finite distance, trace the rays from three object points $\overline{Y}_o = (\overline{Y}_o)_{max.}$, $\overline{Y}_o = 0.7\ (\overline{Y}_o)_{max.}$ and $\overline{Y}_o = 0.5\ (\overline{Y}_o)_{max.}$. For each obliquity, trace at least five meridional rays to enter the entrance pupil at uniform intervals ranging from $Y_1 = (Y_1)_{max.}$ to $Y_1 = -(Y_1)_{max.}$.

(3) For each obliquity, trace three skew rays with coordinates in the entrance pupil as follows:

$(X_1)_{max.}$, $Y_1 = 0$
$0.7\ (X_1)_{max.}$, $Y_1 = 0$
$0.5\ (X_1)_{max.}$, $Y_1 = 0$

$(X_1)_{max.} = (Y_1)_{max.}$ since the entrance pupil is assumed to be a circle.

(4) Repeat steps 2 and 3 for F and C light.

9.2.5.3 The data from the ray tracing may be plotted as illustrated in Figure 8.5, and Figures 8.7 through 8.10. In practice, this data is plotted on a single diagram usually leaving out the plots shown in Figures 8.8b and 8.9 . A plot of this type is shown in Figure 9.1. In making these plots, it is advisable to use the same scale for all the plots of Y_k and X_k . At first it might appear that lenses of different focal lengths should be plotted using different scales. Actually, for most applications, the scale shown in Figure 9.1 represents the size of images used most frequently. Therefore, it simplifies plotting and helps one to assess rapidly a lens if these plots are made on this standard scale. Notice that 0.01 division on the vertical scale corresponds to 1 cm. (But this has been reduced to 0.86 cm in reproduction.) If the lens is calculated in centimeters, then 1 cm on the vertical scale of the graph corresponds to 100 microns. If the lens is calculated in inches, the 0.01 division should be replaced by 0.004, so that again 1 cm indicates a 100 micron image. If it turns out that the aberrations are so large they cannot be plotted on this scale, they are so large that they probably are not worth plotting.

9.2.6 Step 6 - Adjusting third order aberrations. Usually one attempts to make the curves in Figure 9.1 as flat as possible. In a perfect lens the curves would be horizontal straight lines. In most cases this can not be achieved, even to practical limits. The usual curves look more like the ones shown in Figure 9-1. Take for example the curves shown for the image point at 1.76. The meridional rays are focused within a strip 0.012 wide. The skew rays are confined within a strip 0.016 wide. One can say with fair assurance that the complete image is confined to an area 0.012 by 0.016. Since the meridional ray plot

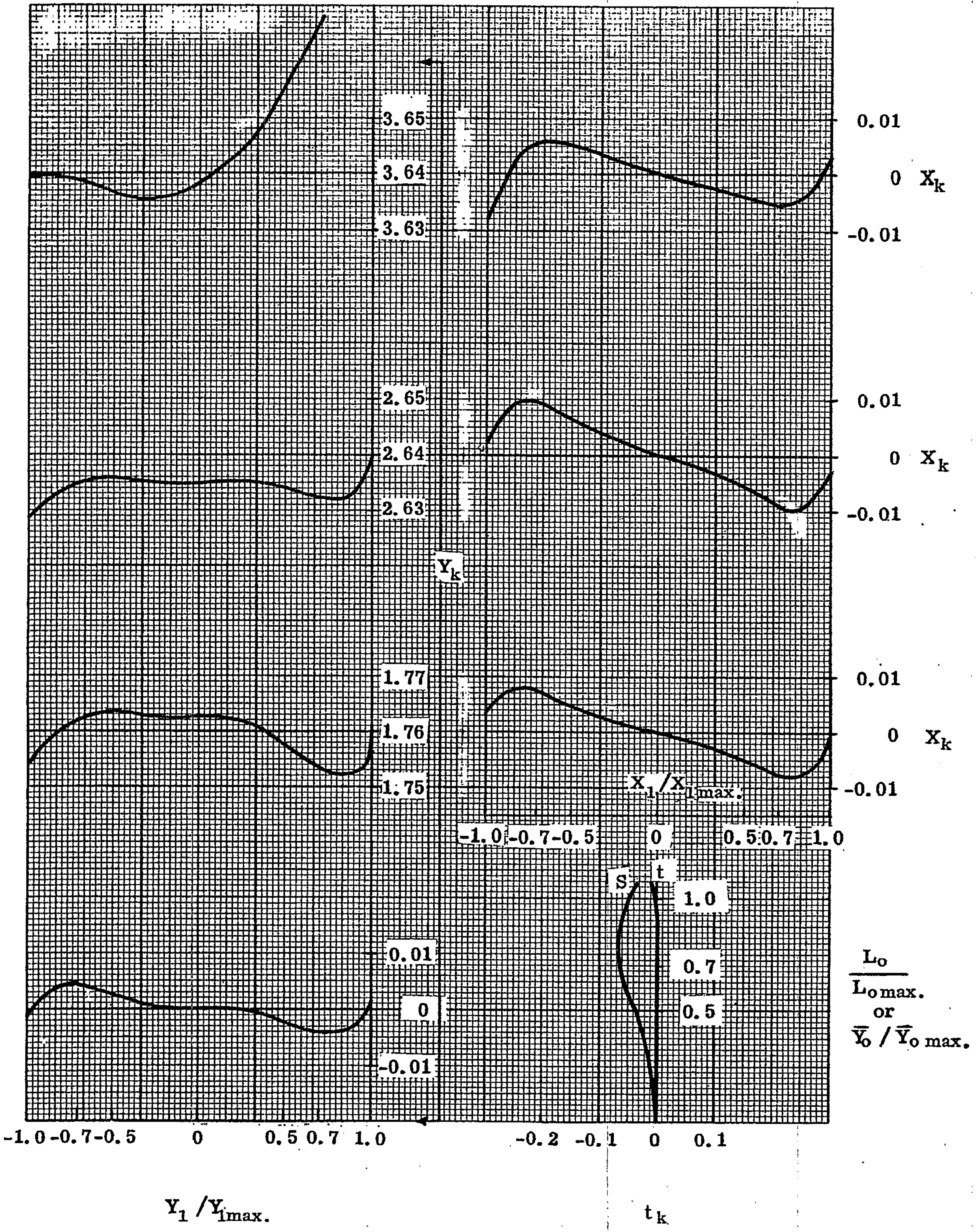


Figure 9.1- Sample plot of selected ray trace data.

shows a region where the curve is flat and horizontal, one would expect to get some concentration of energy towards the center of the spot. When one begins adjusting a design it is usually possible to tell from these curves what is needed to improve the energy concentration. For example, the basic difficulty with the design represented by Figure 9.1 is that the Petzval sum is too negative. This is the reason the skew curves are so far from the horizontal. One can also see that the image at a height of 3.64 will be poor because of the over-corrected spherical aberration in the upper meridional rays. These defects might suggest to the designer that he should try to find a solution with less negative Petzval curvature and introduce more negative third order spherical aberration. If so he would then return to Step 4 in paragraph 9.2.4 and solve for new third order aberrations, and repeat Step 5. Several alternate solutions may therefore evolve, but eventually it will be necessary to evaluate the energy concentration by proceeding to Step 7.

9.2.7 Step 7 - Calculation of spot diagrams and energy distributions. The energy distribution curves should be computed as described on page 8.1. Usually it is advisable to compute the energy distribution curves for a field point on the axis, for one half-way out in the field and for one at the edge of the field. Strictly speaking, one should also compute curves for two or three wavelengths, but this takes a great deal of computing and usually is not necessary for the average problem.

9.2.8 Step 8 - Image evaluation.

9.2.8.1 Once the designer has computed the energy distributions in several images in the field he is able to compare these with the design requirements. Seldom can one achieve the required results in the first system analyzed. The designer must then decide whether to continue with this design or to shift over to another type of lens. If he shifts over to another lens type he may then return to Step 1. If he decides to stick with the present lens type, he must decide whether to continue trying to meet the original specifications or whether to seek to modify the specifications and provide an alternative compromise solution. Usually the modern design problems end up with a give and take solution. The designer must therefore completely understand how the lens will perform, and be able to show what can be achieved by making variations in the original specifications. This means he may have to carry several designs up to the energy distribution curves in Step 7 in order to make a wise decision. It is imperative therefore that he devise ways to quickly evaluate the design.

9.2.8.2 The energy distribution curves of Step 7 may be used to check the image quality. This is a satisfactory method for many optical systems, but if the image quality is high one must consider the calculation of diffraction effects. As a general rule, one does not need to worry about diffraction effects if the wavefront departs from a perfect sphere by more than two to five wavelengths. (A method of computing this departure from ray trace data is described by H. H. Hopkins.*) There are several criteria one can apply to gauge the influence of diffraction, but this is a subject in itself. (See Sections 16, 25, 26.) However, a designer should be familiar with the wavefront tolerances suggested by Conrady.**

9.2.8.3 One must remember that it is impossible to concentrate the energy in an image into a smaller spot size than predicted by diffraction. In Figure 9.2 a plot of energy distribution is shown for a perfect lens. The abscissa Z is the following:

$$Z = \frac{\pi \; Y_e d}{\lambda \; \ell'} ,$$

where

Y_e is the radius of the exit pupil

d is the diameter of the image spot

λ is the wave length of light

ℓ' is the distance from the exit pupil to the image plane which is located at the perfect focus.

The first dark ring occurs at a value of Z equal to 3.83. This has a spot diameter

$$d = \left(\frac{3.83\,\lambda}{\pi}\right)\left(\frac{\ell'}{Y_e}\right) = 1.22 \; \frac{\lambda\,\ell'}{Y_e}$$

---

* H. H. Hopkins, Wave Theory of Aberrations (Oxford University Press, London, 1950) pp. 21-23.

**A. E. Conrady, Applied Optics and Optical Design, Part 1 (Oxford University Press, New York, 1943) pp. 126-141. See also Part I, 2nd ed. (Dover, New York, 1957) pp. 126-141, and Part II (Dover, 1960) pp. 626-639.

It is always a good idea to plot this curve on the same graph with the energy distribution curves computed for the actual lens. If the geometrical energy distribution curves lie to the left of the diffraction curve one knows that the light will not concentrate as well as the geometrical distribution curves indicate. The actual distribution curve will be inclined to follow the diffraction curve. Quite often the geometrical energy distribution curve will cross the diffraction image curve as shown in Figure 9.3 One can then estimate the energy concentration by using the formula

$$Z_{G+D} = \sqrt{Z_{G}^{2} + Z_{D}^{2}} \;.$$

Where Z_G ~ geometrical spot diameter

Z_D ~ diffraction spot diameter of a perfect aperture

Z_{G+D} ~ estimated spot diameter.

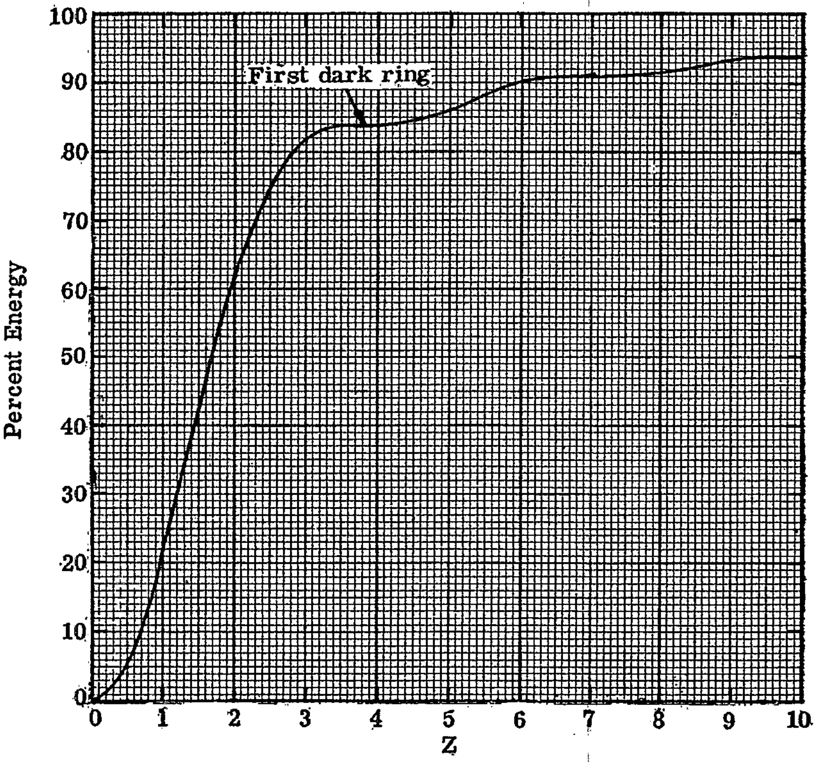


Figure 9.2 - Energy distribution for a perfect lens.

9.2.8.4 Some designers object to the energy distribution method for image evaluation because it does not take into account the orientation of the energy distribution. For example, if there is astigmatism the energy will be concentrated in a line image. The fact that the image of a point is a line might actually be favorable in some types of optical systems. For example, if the image is scanned by a slit one could certainly use this to advantage. For most optical systems however the circular energy distribution curves are adequate.

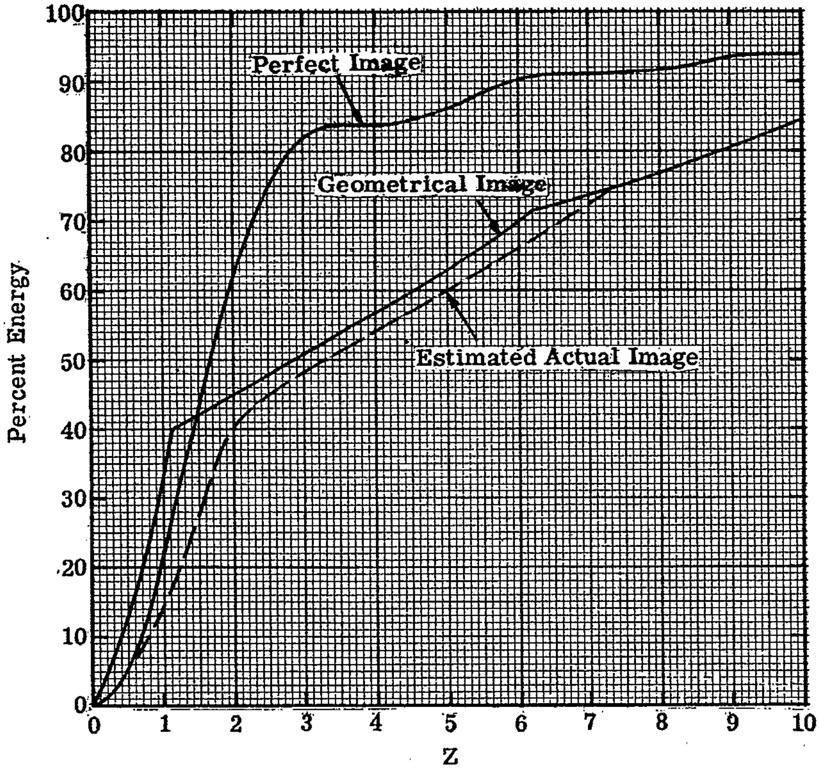


Figure 9.3 - Energy distribution curves.

9.2.8.5 The most modern method for evaluating images is to compute the optical transfer function (often called the sine wave frequency response) for the image. This can be done by performing a Fourier transform of the energy distribution in the image of a point source or a line source. Figure 9.4 shows a series of energy distribution curves. Figure 9.5 shows the corresponding modulation transfer curves. The modulation transfer function is the modulus of the complex optical transfer function. In Figure 9.4 all the curves are normalized to a maximum spot diameter of 10 mm. In Figure 9.5 the frequency is given in lines/mm. These spot diameters may of course be scaled to any other size. For example, suppose the maximum spot diameter is 100μ. Then the frequency scale should be multiplied by 100. One can multiply the modulation transfer function of a lens by the function for a detector to obtain the overall function for the entire optical system. Finally one can estimate the cutoff frequency at some particular response. A review of this approved image evaluation method may be found in Sections 26.2, 26.3 and 26.4, and in the article by Perrin ("Methods of Appraising Photographic Systems," J. Soc. Motion Picture and Television Eng., 69, 151-156, 239-249 (1960).

9.2.8.6 The problem of image evaluation is so involved that actually a designer is always forced to refer to some system which is known. Before attempting to improve a new system, a designer should try to do the following:

(1) Find out what systems have already been designed for conditions as nearly identical as possible with those specifying the new system.
(2) Evaluate the energy distribution of the nearest equivalent system.
(3) Compare the energy distribution in the new design with that of the closest equivalent to determine if improvement has been made.

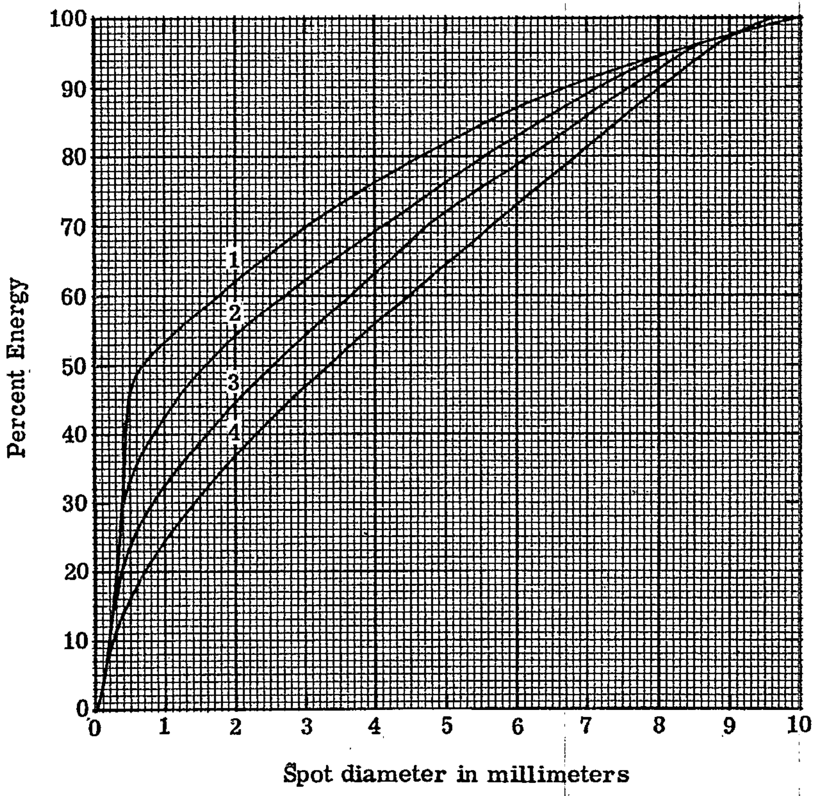


Figure 9.4 - A series of energy distribution curves.

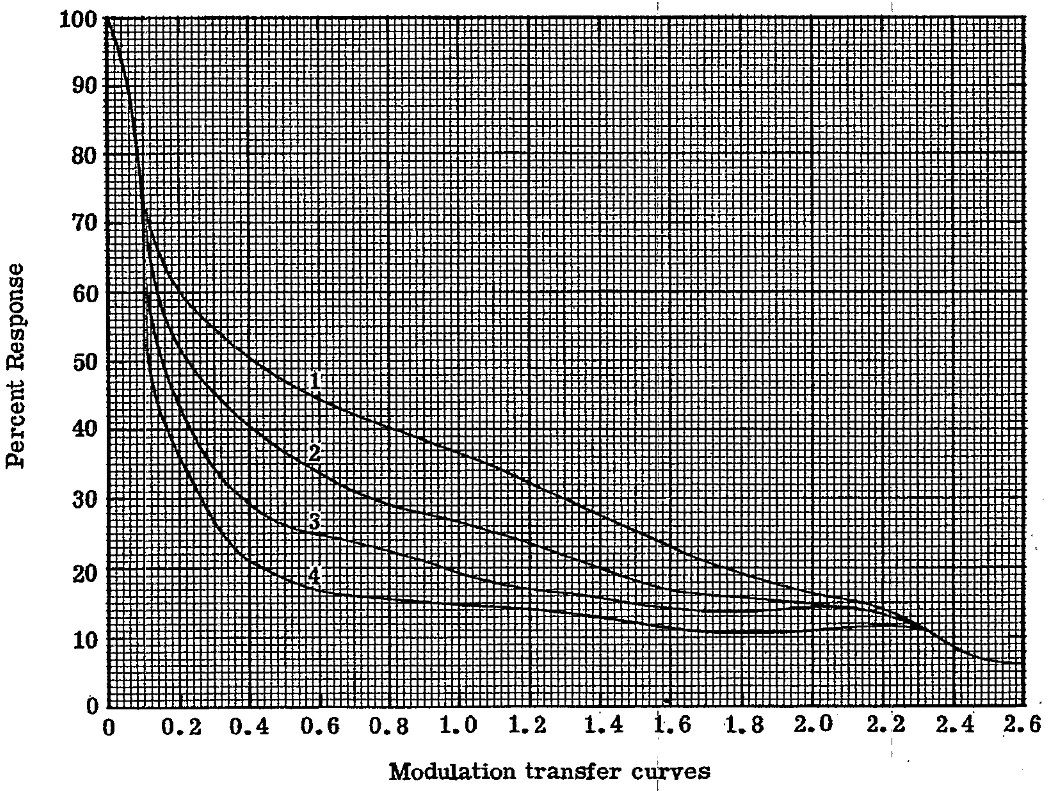


Figure 9.5 -

9.3 SUMMARY OF EQUATIONS USED IN THE CALCULATION OF THIRD ORDER ABERRATIONS

9.3.1 Paraxial ray trace equations.

$y = y_{-1} + (t_{-1}/n_{-1})(n_{-1}u_{-1})$ 5-(56)

$nu = n_{-1}u_{-1} + y(n_{-1} - n)c$ 5-(57)

Alternate equations,

$y = y_{-1} + t_{-1}u_{-1}$ 5-(56)

$u = u_{-1} + i\left(\frac{n_{-1}}{n} - 1\right)$ 6-(3)

$i = yc + u_{-1}$ 6-(4)

$\Phi = \bar{y}(nu) - y(n\bar{u}) = \bar{y}(n_{-1}u_{-1}) - y(n_{-1}\bar{u}_{-1})$ 6-(6)

9.3.2 Chromatic contribution formulae.

$a = -yn_{-1}\, i\left(\frac{dn}{n} - \frac{dn_{-1}}{n_{-1}}\right)$ 6-(34)

$b = -yn_{-1}\,\bar{i}\left(\frac{dn}{n} - \frac{dn_{-1}}{n_{-1}}\right)$ 6-(35)

9.3.3 Third order surface contributions.

$S = yn_{-1}\left(\frac{n_{-1}}{n} - 1\right)(u + i)$ 8-(5)

$\bar{S} = \bar{y}n_{-1}\left(\frac{n_{-1}}{n} - 1\right)(\bar{u} + \bar{i})$ 8-(13) footnote

$B = Si^2$ 8-(4)

$F = Si\bar{i}$ 8-(11)

$C = S\bar{i}^2$ 8-(12)

$E = \bar{S}i\bar{i} + \Phi(\bar{u}_{-1}^2 - \bar{u}^2)$ 8-(13)

$P = \frac{c(n_{-1} - n)}{n_{-1}\, n}$ 8-(14)

For an aspheric surface with a fourth order coefficient of e,

$B = 8(n_{-1} - n)ey^4$ 8-(4a)

$F = B\bar{y}/y$ 8-(11a)

$C = B(\bar{y}/y)^2$ 8-(12a)

$E = B(\bar{y}/y)^3$ 8-(13a)

9.3.4 Stop shift equations.

$\Sigma B^* = \Sigma B$ 8-(18)

$\Sigma F^* = Q\Sigma B + \Sigma F$ 8-(19)

$\Sigma C^* = Q^2\Sigma B + 2Q\Sigma F + \Sigma C$ 8-(20)

$\Sigma E^* = Q^3\Sigma B + 3Q^2\Sigma F + Q\Sigma(3C + P\Phi^2) + \Sigma E$ 8-(21)

$\Sigma P^* = \Sigma P$ 8-(21a)

$a^* = a$ 8-(22)

$b^* = Qa + b$ 8-(23)

10 AN APPLICATION OF THE METHOD OF LENS DESIGN

10.1 STEP ONE - SELECTING THE LENS TYPE

10.1.1 The Taylor triplet. In order to illustrate the procedure described in Section 9, we shall now work through the design of a particular type of lens. The lens selected for illustration is the famous triplet, often referred to as the Taylor triplet. It is named after H. Dennis Taylor who first described how he was able to correct astigmatism and field curvature by using three air spaced lenses. His system consisted of a negative lens between two positive lenses.

10.1.2 Reasons for selection. The triplet lens system is a fundamental type, for there are enough degrees of freedom to specify the first order properties and to control all the first and third order aberrations. First order properties includes the focal length and the optical invariant. First order aberrations are axial and lateral color, and Petzval curvature. Third order aberrations are spherical aberration, coma, astigmatism, and distortion. This lens illustrates most of the problems encountered in the design of any optical system; many of the other types of lenses are merely derivatives of the basic triplet. The triplet has been used extensively in optics; there are probably more such objectives used in photographic instruments than any other type of lens. In describing this design procedure it is hoped that the logical design of an objective can be illustrated; at the same time it will be shown how exceedingly involved the design of a lens can become if it is necessary to arrive at an optimum solution.

10.1.3 Arrangement and notation. The lens arrangement for the triplet objective is shown in Figure 10.1 with the notation to be used in the following discussion. The lens is to work with an object at infinity and have a focal length of 10. It will be color corrected for F and C light. The individual lenses are shown as thick lenses but in the first stages of the design these lenses are assumed to be thin. By selecting this type of lens (the Taylor triplet), step 1 in the design procedure has been completed. The application of the method of design will therefore continue with step 2.

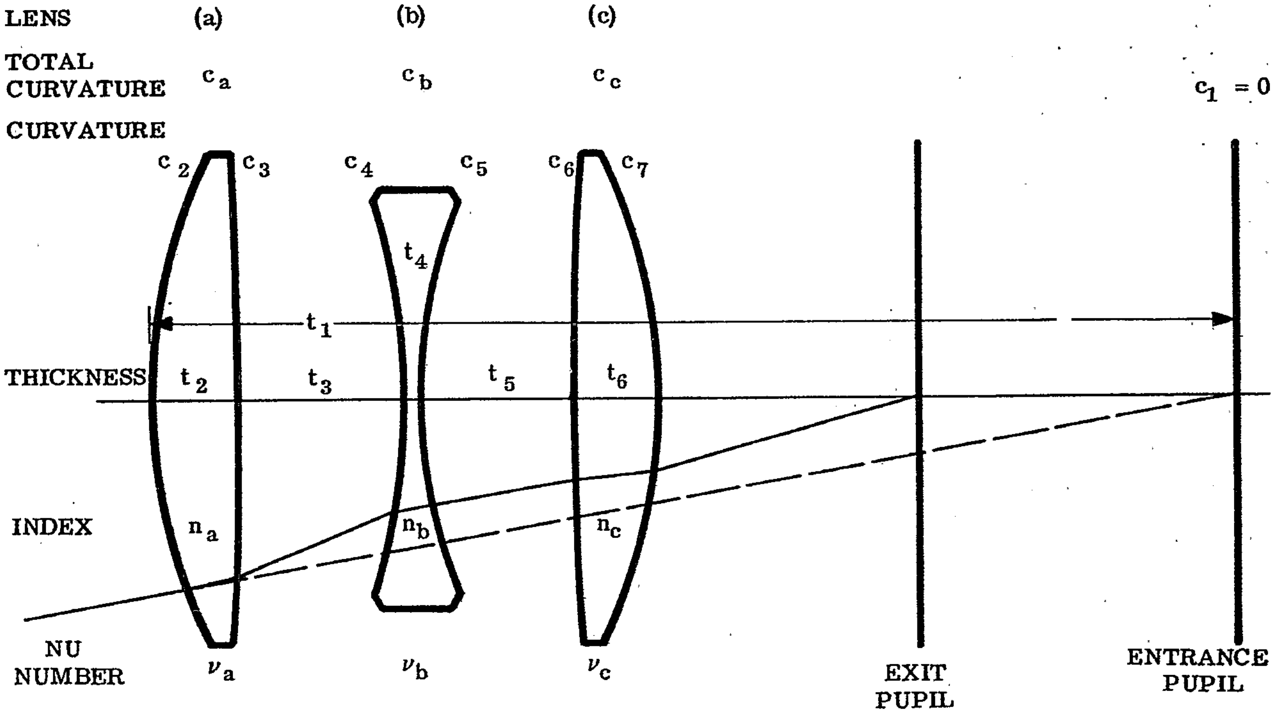


Figure 10.1 - A triplet objective used to illustrate design procedure.

10.2 STEP TWO - THE FIRST ORDER THIN LENS SOLUTION

10.2.1 Power and spacing.

10.2.1.1 The first problem is to decide on the power and spacing of the elements. This is a lens system composed of three thin elements; therefore we immediately set up a thin lens table of the type shown in Table 6.14, and start to fill in the known quantities as shown in Table 10.1. At this stage nothing is known about the design, except that the object is to be at infinity ($u_o = 0$) and the focal length should be 10. Since y_1 may have any value, we choose 1 for convenience. From Equation 6-(13) if $f' = 10$ and $y_1 = 1$ then $u_{k-1} = -0.1$. The computing table appears now as shown in Table 10.2.

SURFACE NO.	Entrance Pupil 1	Lens (a) 2, 3	Lens (b) 4, 5	Lens (c) 6, 7	Focal Plane k
$-\phi$ t	0				0
y u 0					0
$\bar{y}$ $\bar{u}$	0				

Table 10.1- Computing table 1 - quantities known at start of procedure.

SURFACE NO.	Entrance Pupil 1	Lens (a) 2, 3	Lens (b) 4, 5	Lens (c) 6, 7	Focal Plane k
$-\phi$ t	0				0
y u 0	1 0	1		 -0.1	0
$\bar{y}$ $\bar{u}$	0				

Table 10.2- Computing table 2 - first order assumptions added.

10.2.1.2 In order to specify the lateral color for F and C light, the conditions given in Equation 6-(41) must be fulfilled. Thus,

$$Tch_{F-C} = \frac{1}{n_{k-1}\ u_{k-1}} \left[\frac{y_a\ \bar{y}_a\ \phi_a}{\nu_a} + \frac{y_b\ \bar{y}_b\ \phi_b}{\nu_b} + \frac{y_c\ \bar{y}_c\ \phi_c}{\nu_c} \right] .$$

If Tch_{F-C} is to be zero, then

$$\frac{y_a\ \bar{y}_a\ \phi_a}{\nu_a} + \frac{y_b\ \bar{y}_b\ \phi_b}{\nu_b} + \frac{y_c\ \bar{y}_c\ \phi_c}{\nu_c} = 0 .$$

If we assume the condition that the chief ray shall pass through the center of lens (b) (not as shown in Figure 10.1), then $\bar{y}_b = 0$, and

$$\frac{y_a \bar{y}_a \phi_a}{\nu_a} = -\frac{y_c \bar{y}_c \phi_c}{\nu_c} .$$

Equation 6-(24) shows that in a thin lens $y\phi = (u_{-1} - u)$, which is the angular deviation that the axial ray experiences as it passes through the lens. If R is defined as

$$R = \frac{y_a \phi_a}{y_c \phi_c} ,$$

the condition for zero lateral color is then

$$\frac{\bar{y}_a}{\bar{y}_c} = -\frac{1}{R} \frac{\nu_a}{\nu_c} .$$

Since the chief ray passes through the center of the thin negative (b) lens, it is undeviated. Therefore $\bar{y}_b \phi_b = 0$, and $(\bar{u}_{-1} - \bar{u})_b = \bar{u}_4 - \bar{u}_5 = 0$. Then

$$\frac{t_3}{t_5} = -\frac{\bar{y}_a}{\bar{y}_c} = \frac{1}{R} \frac{\nu_a}{\nu_c} .$$

10.2.1.3 Up to this point, no decision has had to be made with respect to the type of glass. At this point it is necessary to decide on ν_a / ν_c. Any ratio may be used, but up until the present no one has been able to prove any advantage to a ratio other than 1. If the same glass is used for both lens (a) and lens (c), then $\nu_a / \nu_c = 1$. This choice has the practical advantage that the lens maker does not have to worry about two different glasses for the positive lenses. (Any designer who uses two elements that look alike but are of slightly different index and/or dispersion can fully expect to find the elements switched in the prototype!) With no positive evidence indicating that ν_a / ν_c needs to be other than 1, the design will proceed with glass (a) and glass (c) the same. Then it follows that

$$t_3 = \frac{1}{R} t_5 .$$

10.2.1.4 Next, it is necessary to choose a value of R. Such a value may be selected for any number of reasons. For each value of R there are many solutions (designs). In the following study an attempt will be made to show how the choice of R affects the design, but in order to proceed with the numerical example it is necessary to assume a value of R. Later (Paragraph 10.3.2.3) it will be shown that R should be near 1. This means that the (a) lens will bend the axial ray through the same angle as does the (c) lens. It follows then, that if a value can be assigned to u_3, the angle the axial ray makes with the axis after emerging from the (a) lens, then the angle u_5 is determined. (This means if u_3 is assigned, then all the angles the axial ray makes with the axis are known). For example, if u_3 is made -0.20, then it follows immediately that $u_5 = 0.10$ because $u_{k-1} = u_7 = -0.1$.

10.2.1.5 The computing table may now be filled out as shown in Table 10.3. It is still not possible to compute ϕ_a, ϕ_b, ϕ_c, and complete the table. At this point it is necessary to make another guess. Let the guess be that the space t_3 will be 1; then t_5 must also be 1. Now the system is completed and ϕ_a, ϕ_b, and ϕ_c are determined using Equations 6-(23) and 6-(24). The values are shown in Table 10.4, which is filled out completely. To trace the chief ray any angle may be assumed for it while it passes through the (b) lens. In the example, $\bar{u}_b = 0.5$ was used.

10.2.2 Glass types.

10.2.2.1 So far the only decision on glass is that (a) = (b). Now we must specify the type of glass to use for (a) and (c), and for (b). The glass types are chosen now in order to specify both axial color and Petzval curvature. When the glasses are chosen, TAch is calculated by Equation 6-(40). This calculation is illustrated in Table 6.13. The Petzval sum, ΣP, may be calculated for each lens from Equation 8-(28) and summed for the (a), (b), and (c) lenses.

10.2.2.2 The choice of glass is a critical part of the design of a triplet. It is hoped that this will be demonstrated in the following study, but in order to show this, the glasses will be picked from experience. The following glasses will be used:

	n_D	ν
Lens (a)	1.620	60.3
Lens (b)	1.617	36.6
Lens (c)	1.620	60.3

SURFACE		Entrance Pupil 1		Lens (a) 2, 3		Lens (b) 4, 5		Lens (c) 6, 7		Focal Plane k
$-\phi$		0								0
t					1		1			
y		1		1						0
u	0		0		-0.2		0.1		-0.1	
$\bar{y}$		0				0				
$\bar{u}$										

Table 10.3- Computing table 3-quantities for zero lateral color, $\nu_a/\nu_c = 1$, and R = 1 , added.

SURFACE		Entrance Pupil 1		Lens (a) 2, 3		(Lens (b) 4, 5		Lens (c) 6, 7		Focal Plane k
$-\phi$		0		-0.2		0.375		-0.222		0
t			-1.25		1		1		9	
y		1		1		0.8		0.9		0
u	0		0		-0.2		0.1		-0.1	
$\bar{y}$		0		-0.5		0		0.5		4.0
$\bar{u}$			0.4		0.5		0.5		0.389	

$f' = 10$

Table 10.4- Computing table 4 - assignment of quantities completed.

10.2.2.3 With this glass type data it is now possible to compute TAch and ΣP. The calculations for the sample are included in Table 10.5.

10.2.2.4 The values of TAch and ΣP are plotted in Figure 10.2. The dot with a surrounding square, ⊡ , indicates where the solution should be for TAch = 0 and ΣP = -0.03. At this point it will be necessary to merely accept the fact that ΣP is set at -0.03. (A negative value of ΣP indicates a negative value for the Petzval curvature. In the case of the triplet example here considered, the field is concave toward the lens and is referred to as an inward curving field.) The next step is to assume a new value of t_3 . For example, suppose we pick a value of 1.25, and repeat the process to arrive at a new value for TAch and ΣP. This point is also plotted in Figure 10.2. Next, set u_3 = 0.18 and repeat the process with t_3 = 1.0 and 1.25. The procedure for finding the values of u_3 and t_3 which will provide a solution follows obviously. With a small amount of practice one can box in a design in this manner in very short order. This is an iterative procedure which can also be programmed for automatic correction on a computer. The graphs obtained in this manner are extremely useful for visualizing how to readjust the angles in the lens after the thickness has been added (step 4).

SURFACE		Lens (a)		Lens (b)		Lens (c)		Focal Plane
$-\phi$		-0.2		0.375		-0.222		0
t								
y		1		0.8		0.9		0
u	0		-0.2		0.1		-0.1	
$\bar{y}$								
$\bar{u}$								
ν		60.3		36.6		60.3		
$-\phi y^2/\nu$		-0.00332		0.00656		-0.00299		$\Sigma a = 0.00026$
$-\frac{\phi}{n}$		-0.12346		0.23191		-0.13717		$\Sigma P = -0.02872$

TAch = 0.0026

Table 10.5 -Computing table 5 - calculation of TAch and ΣP.

10.2.2.5 This boxing in procedure is recommended for the preliminary set-up using thin lenses in designing a triplet. The procedure works equally well for more complicated lenses, and it provides the designer a graphical picture of how the variables affect the system. For those who prefer to manipulate algebraic equations a procedure similar to the above can be worked out to provide equations to be solved. Existing literature is adequately filled with methods of this type. A few of the well known papers are:

(1) Berek, M., Grundlagen der Praktischen Optik, Berlin, 123-130, (1930).

(2) Stephens, R. E. J. Opt. Soc. Am. 38, 1032 – 1039, (1948).

(3) Lessing, N., J. Opt. Soc. Am. 48, 558-562 (1958).

(4) Cruikshank, F. D. Rev. D'Optik 35, 292-299, (1956).

(5) Cruikshank, F. D. Australian J. Physics 11, 41-54, (1958).

A series of solutions for triplets with different types of glass has been worked out. The significant data for these systems are included in Table 10.6, sheets 1, 2 and 3. The glasses used in this study are shown plotted in Figure 10.3 on an n_D versus ν plot, which is used extensively by lens designers. The numbers alongside each point indicate the system number. The table includes calculations for $R = 1, 0.5$ and 2. One solution was calculated, for each set of glasses, using a target value of $\Sigma P = -0.03$. Notice that there are examples where $(\nu_a - \nu_b)$ is constant but $(n_a - n_b)$ changes.

10.2.3 Summary of thin lens first order study contained in Table 10.6.

(1) As $\nu_a - \nu_b$ is increased, the system length, T, always increases.

(2) $R = 1$ systems are always shorter than systems with $R = 2.0$ or $R = 0.5$.

(3) Changing ΣP from -0.03 to -0.02 shortens the system.

(4) Changing the index of the crown and flint elements, while maintaining the ν difference, has little effect on the overall length T.

(5) As one would expect, the higher the index of the positive elements, the lower the power of all the elements.

(6) Solutions for $R = 2$ and $R = 0.5$ are essentially inverted solutions. t_3 and t_5 are almost exactly interchanged. Also, ϕ_a is changed by the ratio of $1/R$.

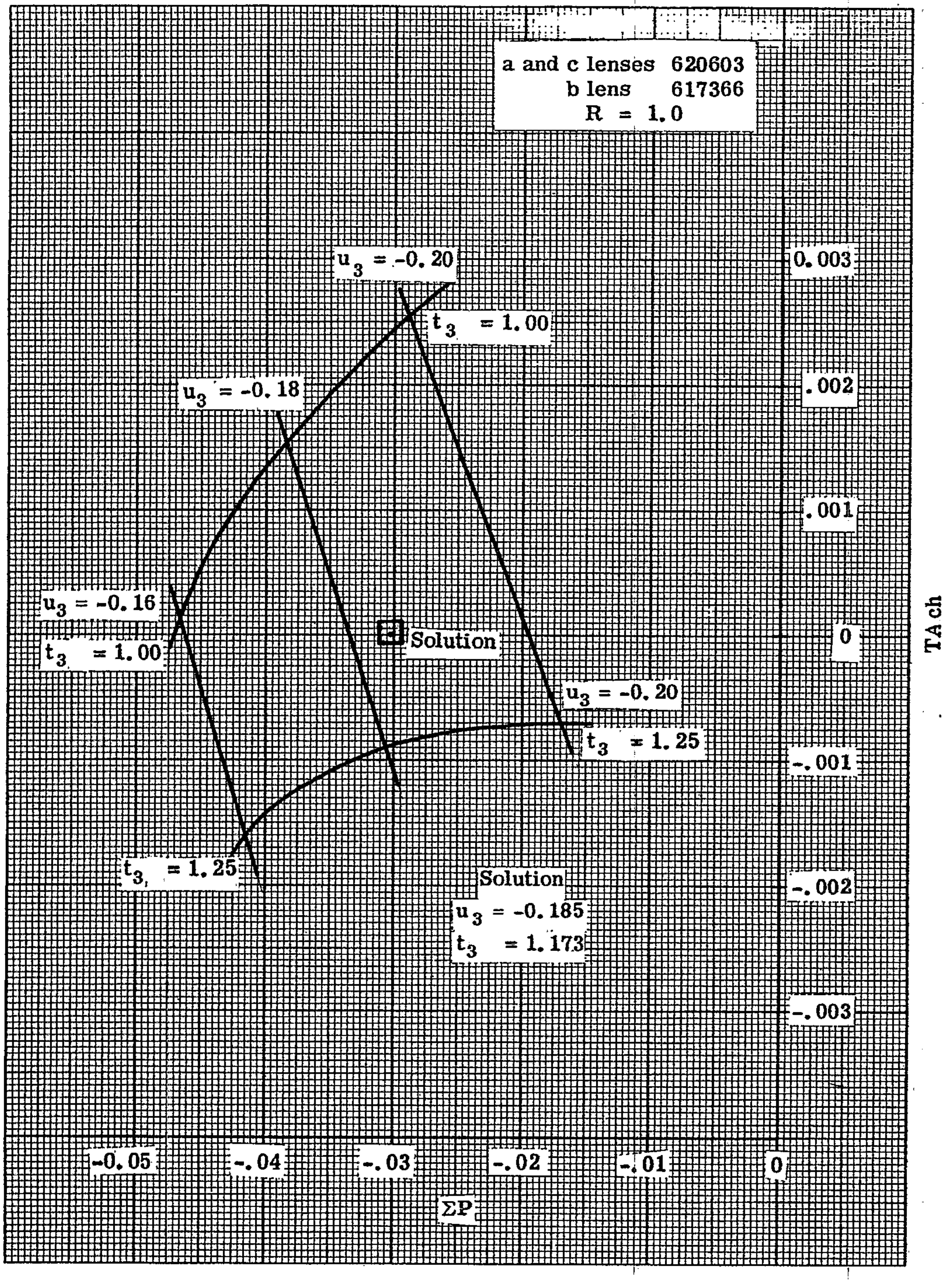


Figure 10.2 - Diagram used to find a thin lens first order solution.

System No.	$R = \frac{t_5}{t_3}$	Glass (a and c) (lenses)	Glass (b lens)	$n_a - n_b$	$\nu_a - \nu_b$	ϕ_a	t_3	t_5	$T = t_3 + t_5$
1A	1	511635	596397	-.0846	23.8	0.220	1.062	1.062	2.124
1B	2	"	"	"	"	0.296	0.795	1.590	2.385
1C	0.5	"	"	"	"	0.146	1.583	0.792	2.375
2A	1	511635	617366	-.1060	26.9	0.195	1.442	1.442	2.884
2B	2	"	"	"	"	0.264	1.075	2.150	3.225
2C	0.5	"	"	"	"	0.130	2.138	1.069	3.207
3A	1	511635	649338	-.1380	29.7	0.180	1.850	1.850	3.700
3B	2	"	"	"	"	0.244	1.370	2.740	4.110
3C	0.5	"	"	"	"	0.1185	2.736	1.368	4.104
4A	1	511635	657366	-.1460	26.9	0.205	1.419	1.419	2.838
4B	2	"	"	"	"	0.275	1.066	2.132	3.198
4C	0.5	"	"	"	"	0.135	2.128	1.064	3.192
5A	1	541599	596397	-.0546	20.2	0.236	0.817	0.817	1.634
5B	2	"	"	"	"	0.317	0.615	1.230	1.845
5C	0.5	"	"	"	"	0.158	1.229	0.615	1.844
6A	1	541599	617366	-.0760	23.3	0.207	1.161	1.161	2.322
6B	2	"	"	"	"	0.278	0.874	1.748	2.622
6C	0.5	"	"	"	"	0.138	1.750	0.875	2.625
*6AA	1	541599	617366	-.0760	23.3	0.222	1.157	1.157	2.314
*6BB	2	"	"	"	"	0.298	0.867	1.734	2.601
*6CC	0.5	"	"	"	"	0.148	1.732	0.866	2.598
7A	1	541599	649338	-.1080	26.1	0.189	1.545	1.545	3.090
7B	2	"	"	"	"	0.255	1.155	2.310	3.465
7C	0.5	"	"	"	"	0.1245	2.300	1.150	3.450
8A	1	541599	657366	-.1160	23.3	0.218	1.154	1.154	2.308
8B	2	"	"	"	"	0.293	0.868	1.736	2.604
8C	0.5	"	"	"	"	0.147	1.736	0.868	2.604
9A	1	541599	689309	-.1480	29.0	0.172	2.005	2.005	4.010
9B	2	"	"	"	"	0.236	1.500	3.000	4.500
9C	0.5	"	"	"	"	0.113	3.000	1.500	4.500
10A	1	588612	596397	-.0076	21.5	0.211	0.880	0.880	1.760
10B	2	"	"	"	"	0.284	0.660	1.320	1.980
10C	0.5	"	"	"	"	0.141	1.330	0.665	1.995

Table 10.6- Thin lens triplet (first order solution), Sheet 1 of 5

System No.	$R = \frac{t_3}{t_5}$	Glass (a and c) (lenses)	Glass (b lens)	$n_a - n_b$	$\nu_a - \nu_b$	ϕ_a	t_3	t_5	$T = t_3 + t_5$
11A	1	588612	617366	-.0290	24.6	0.188	1.259	1.259	2.518
11B	2	"	"	"	"	0.253	0.947	1.894	2.841
11C	0.5	"	"	"	"	0.127	1.894	0.947	2.841
12A	1	588612	649338	-.0610	27.4	0.173	1.658	1.658	3.316
12B	2	"	"	"	"	0.235	1.246	2.492	3.738
12C	0.5	"	"	"	"	0.118	2.492	1.246	3.738
13A	1	588612	657366	-.0690	24.6	0.195	1.264	1.264	2.528
13B	2	"	"	"	"	0.264	0.950	1.900	2.850
13C	0.5	"	"	"	"	0.132	1.900	0.950	2.850
14A	1	611588	617366	-.0060	22.2	0.195	1.080	1.080	2.160
14B	2	"	"	"	"	0.261	0.790	1.580	2.370
14C	0.5	"	"	"	"	0.129	1.570	0.785	2.355
15A	1	620603	596397	.0244	20.6	0.207	0.810	0.810	1.620
15B	2	"	"	"	"	0.280	0.614	1.228	1.842
15C	0.5	"	"	"	"	0.140	1.228	0.614	1.842
16A	1	620603	617366	.0030	23.7	0.185	1.173	1.173	2.346
16B	2	"	"	"	"	0.248	0.882	1.764	2.646
16C	0.5	"	"	"	"	0.124	1.764	0.882	2.646
17A	1	620603	621362	-.0010	24.1	0.184	1.240	1.240	2.480
17B	2	"	"	"	"	0.246	0.930	1.860	2.790
17C	0.5	"	"	"	"	0.121	1.830	0.915	2.745
18A	1	620603	649338	-.0290	26.5	0.170	1.600	1.600	3.200
18B	2	"	"	"	"	0.231	1.180	2.360	3.540
18C	0.5	"	"	"	"	0.113	2.353	1.177	3.530
19A	1	620603	657366	-.0370	23.7	0.193	1.188	1.188	2.376
19B	2	"	"	"	"	0.259	0.895	1.790	2.685
19C	0.5	"	"	"	"	0.130	1.790	0.895	2.685
20A	1	620603	668323	-.0480	28.0	0.163	1.830	1.830	3.660
20B	2	"	"	"	"	0.222	1.372	2.744	4.116
20C	0.5	"	"	"	"	0.111	2.744	1.372	4.116

Table 10.6-Thin lens triplet (first order solution). Sheet 2 of 5

System No.	$R = \frac{t_3}{t_5}$	Glass (a and c) (lenses)	Glass (b lens)	$n_a - n_b$	$\nu_a - \nu_b$	ϕ_a	t_3	t_5	$T = t_3 + t_5$
21A	1	657572	617366	.0400	20.6	0.193	0.911	0.911	1.822
21B	2	"	"	"	"	0.258	0.686	1.372	2.058
21C	0.5	"	"	"	"	0.129	1.372	0.686	2.058
22A	1	657572	649338	.0080	23.4	0.176	1.289	1.289	2.578
22B	2	"	"	"	"	0.238	0.978	1.956	2.934
22C	0.5	"	"	"	"	0.119	1.956	0.978	2.934
23A	1	657572	668323	-.0110	24.9	0.168	1.540	1.540	3.080
23B	2	"	"	"	"	0.228	1.158	2.316	3.474
23C	0.5	"	"	"	"	0.114	2.316	1.158	3.474
24A	1	657572	689309	-.0320	26.3	0.162	1.773	1.773	3.546
24B	2	"	"	"	"	0.221	1.337	2.674	4.011
24C	0.5	"	"	"	"	0.108	2.625	1.313	3.938
25A	1	691548	649338	.0420	21.0	0.180	1.063	1.063	2.126
25B	2	"	"	"	"	0.242	0.805	1.610	2.415
25C	0.5	"	"	"	"	0.121	1.610	0.805	2.415
26A	1	691548	689309	.0020	23.9	0.166	1.530	1.530	3.060
26B	2	"	"	"	"	0.224	1.161	2.322	3.483
26C	0.5	"	"	"	"	0.112	2.322	1.161	3.483
27A	1	691548	720293	-.0290	25.5	0.159	1.836	1.836	3.672
27B	2	"	"	"	"	0.216	1.376	2.752	4.128
27C	0.5	"	"	"	"	0.108	2.752	1.376	4.128
28A	1	720475	689309	.0310	16.6	0.199	0.842	0.842	1.684
28B	2	"	"	"	"	0.266	0.632	1.264	1.896
28C	0.5	"	"	"	"	0.133	1.264	0.632	1.896
29A	1	720475	720293	0000	18.2	0.188	1.082	1.082	2.164
29B	2	"	"	"	"	0.248	0.820	1.640	2.460
29C	0.5	"	"	"	"	0.125	1.620	0.810	2.430
29D	1.5	"	"	"	"	0.228	0.890	1.335	2.225

Table 10.6-Thin lens triplet (first order solution). Sheet 3 of 5

System No.	Vertex of Parabola $_3Y_k$	FD for c_2 at Vertex	Slope of FD Curve	c_2 at Vertex $_3Y_k$	c_4 at Vertex $_3Y_k$	c_6 at Vertex $_3Y_k$	c_a	c_b	c_c
1A	.029	-.008	-.19	.4320	-.2954	.1272	.4305	-.7449	.4817
1B	-.023	.004	-.30	.5150	-.2759	-.0504	.5792	-.7553	.3444
1C	-.030	-.021	-.18	.3750	-.3120	.1960	.2857	-.7381	.6205
5A	.094	-.005	-.11	.4500	-.3126	.1154	.4362	-.7738	.4750
5B	.029	.005	-.16	.5100	-.3015	.0082	.5860	-.7831	.3341
5C	.025	-.014	-.11	.4000	-.3435	.1947	.2921	-.7793	.6223
6A	.014	-.010	-.20	.4000	-.2518	.1209	.3826	-.6699	.4329
6B	-.030	.007	-.27	.4600	-.2466	.0299	.5139	-.6787	.3114
6C	-.032	-.022	-.19	.3400	-.2843	.1698	.2551	-.6709	.5591
*6AA	.033	-.010	-.25	.4200	-.2928	.1322	.4104	-.7502	.4640
*6BB	-.016	.002	-.38	.5050	-.2770	.0534	.5508	-.7583	.3332
*6CC	-.019	-.025	-.20	.3600	-.3128	.1789	.2736	-.7497	.5990
7A	-.021	-.015	-.38	.3650	-.2188	.1115	.3494	-.6050	.4132
7B	-.046	.010	-.50	.4120	-.1650	.0667	.4713	-.6170	.3065
7C	-.050	-.031	-.33	.3000	-.2347	.1521	.2301	-.5905	.5201
8A	-.003	-.010	-.35	.4150	-.2450	.1179	.4030	.6833	.4555
8B	-.054	.002	-.40	.5000	-.2194	.0603	.5416	-.6930	.3277
8C	-.053	-.025	-.24	.3600	-.2710	.1816	.2717	-.6969	.5951
10A	.057	-.005	-.11	.4000	-.2534	.0964	.3588	-.6639	.3935
10B	.016	.003	-.14	.4520	-.2319	.0082	.4830	-.6736	.2782
10C	.010	-.013	-.10	.3450	-.3087	.1490	.2398	-.6675	.5138
11A	.009	-.009	-.22	.3480	-.2170	.0993	.3197	-.5860	.3658
11B	-.020	.006	-.31	.4000	-.2013	.0262	.4303	-.5957	.2654
11C	-.022	-.023	-.19	.3050	-.2558	.1474	.2160	-.5997	.4772
13A	.004	-.010	-.25	.3580	-.2153	.1014	.3316	-.5858	.3796
13B	-.029	.003	-.35	.4250	-.1922	.0386	.4490	-.6014	.2771
13C	-.029	-.024	-.22	.3130	-.2523	.1504	.2245	-.6014	.4961
14A	.027	-.006	-.16	.3500	-.2319	.0828	.3191	-.5954	.3578
14B	-.007	.001	-.18	.4180	-.1769	.0340	.4272	-.5952	.2537
14C	-.012	-.014	-.14	.3130	-.2651	.1346	.2111	-.5833	.4582

Table 10.6- Thin lens triplet (third order solution). Sheet 4 of 5

System No.	Vertex of Parabola $_3Y_k$	FD for c_2 at Vertex	Slope of FD Curve	c_2 at Vertex $_3Y_k$	c_4 at Vertex $_3Y_k$	c_6 at Vertex $_3Y_k$	c_a	c_b	c_c
15A	.059	-.003	-.09	.3800	-.2544	.0655	.3339	-.6334	.3633
15B	.025	.003	-.13	.4350	-.2117	-.0043	.4516	-.6488	.2574
15C	.021	-.010	-.08	.3350	-.3162	.1252	.2258	-.6488	.4812
16A	.016	-.008	-.15	.3380	-.2057	.0780	.2984	-.5589	.3380
16B	-.013	-.002	-.20	.4000	-.1578	.0463	.4000	-.5643	.2428
16C	-.014	-.017	-.13	.2900	-.2563	.1271	.2000	-.5643	.4387
17A	.013	-.009	-.20	.3360	-.2028	.0911	.2968	-.5591	.3388
17B	-.015	.004	-.24	.3960	-.1571	.0548	.3968	-.5617	.2437
17C	-.020	-.018	-.19	.2880	-.2430	.1278	.1952	-.5440	.4296
19A	.014	-.007	-.20	.3400	-.2157	.0832	.3113	-.5648	.3533
19B	-.016	.001	-.28	.4100	-.1674	.0423	.4177	-.5716	.2544
19C	-.019	-.020	-.17	.3050	-.2480	.1390	.2097	-.5753	.4606
21A	.034	-.004	-.10	.3480	-.2170	.0601	.2938	-.5624	.3232
21B	.004	.001	-.18	.4050	-.1541	.0122	.3927	-.5652	.2276
21C	.000	-.010	-.09	.3000	-.2816	.1024	.1963	-.5652	.4216
22A	.005	-.010	-.18	.3200	-.1693	.0873	.2679	-.5022	.3075
22B	-.015	.003	-.27	.3620	-.1468	.0270	.3623	-.5161	.2252
22C	-.017	-.020	-.16	.2750	-.2273	.1192	.1811	-.5161	.4015
25A	.014	-.006	-.13	.3200	-.1792	.0612	.2605	-.4954	.2915
25B	-.007	.001	-.17	.3680	-.1278	.0300	.3502	-.5033	.2087
25C	-.011	-.014	-.13	.2800	-.2410	.0992	.1751	-.5033	.3807
29A	.015	-.006	-.13	.3200	-.1710	.0607	.2611	-.4812	.2928
29B	-.012	.002	-.20	.3650	-.1136	.0130	.3444	-.4742	.2060
29C	-.010	-.016	-.11	.2950	-.2118	.1084	.1736	-.4789	.3778
29D	.008	-.002	-.18	.3520	-.1417	.0261	.3167	-.4879	.2436

Table 10.6-Thin lens triplet (third order solution). Sheet 5 of 5

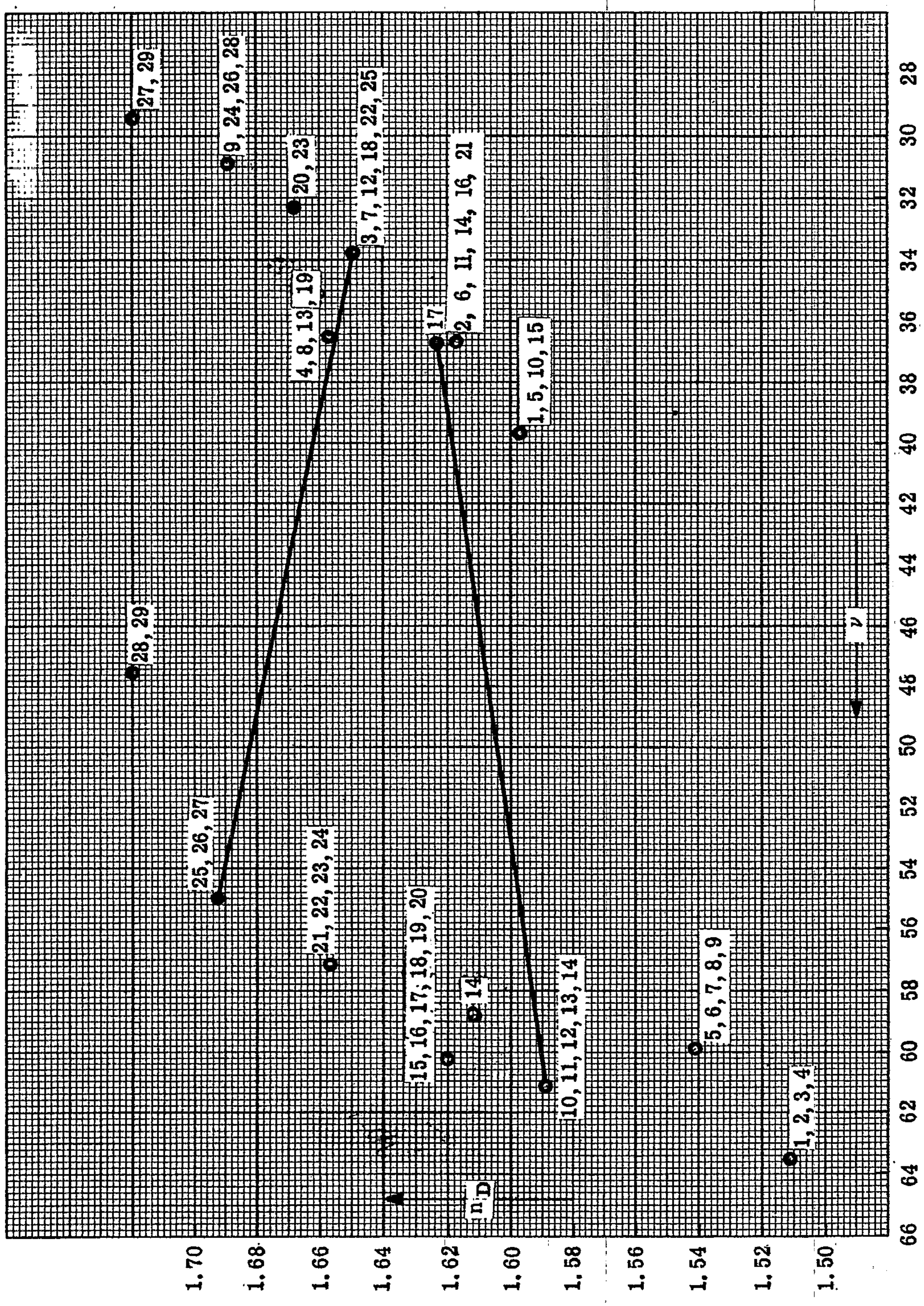


Figure 10.3 - Plot of n_D versus ν for glasses used in triplet study.

10.3 STEP THREE - THE THIRD ORDER THIN LENS SOLUTION

10.3.1 Evaluation of third order coefficients.

10.3.1.1 With the thin lens first order equations worked out so that ϕ_a, ϕ_b, ϕ_c, t_3 and t_5 are known, it is now possible to evaluate the coefficients in Equations 8-(36) through 8-(39) for each lens. In order to simplify the equations, the chief ray is again chosen to pass through the center of the (b) lens. Then the problem is to solve the following equations:

$$B_a^* + B_b + B_c^* = \Sigma B = 0 \tag{1}$$

$$F_a^* + F_b + F_c^* = \Sigma F = 0 \tag{2}$$

$$C_a^* + C_b + C_c^* = \Sigma C = -\frac{1}{3}\,\Sigma P\,\Phi^2 \tag{3}$$

$$E_a^* + 0 + E_b^* = \Sigma E = 0 \tag{4}$$

10.3.1.2 The value of ΣC was not set equal to zero because ΣP is not zero. Instead the value of ΣC is chosen to make $t_{kT} = 0$, as it is defined in Equation 8-(9). For these equations,

$$B_a^* = \alpha_1^* + \alpha_2^*\,c_2 + \alpha_3^*\,c_2^2 \tag{5}$$

$$B_b = \alpha_1 + \alpha_2\,c_4 + \alpha_3\,c_4^2 \qquad \text{Note: The } \alpha\text{'s are} \tag{6}$$

$$B_c^* = \alpha_1^* + \alpha_2^*\,c_6 + \alpha_3^*\,c_6^2 \qquad \text{different for each lens.} \tag{7}$$

$$F_a^* = \beta_1^* + \beta_2^*\,c_2 + \beta_3^*\,c_2^2 \tag{8}$$

$$F_b = \beta_1 + \beta_2\,c_4 \qquad \text{The } \beta\text{'s are different} \tag{9}$$

$$F_c^* = \beta_1^* + \beta_2^*\,c_6 + \beta_3^*\,c_6^2 \qquad \text{for each lens.} \tag{10}$$

$$C_a^* = \gamma_1^* + \gamma_2^*\,c_2 + \gamma_3^*\,c_2^2 \tag{11}$$

$$C_b = -\,\phi_b\,\Phi^2 \qquad \text{The } \gamma\text{'s are different} \tag{12}$$

$$C_c^* = \gamma_1^* + \gamma_2^*\,c_6 + \gamma_3^*\,c_6^2 \qquad \text{for each lens.} \tag{13}$$

$$E_a^* = \delta_1^* + \delta_2^*\,c_2 + \delta_3^*\,c_2^2 \tag{14}$$

$$E_b = 0 \qquad \text{The } \delta\text{'s are different} \tag{15}$$

$$E_c^* = \delta_1^* + \delta_2^*\,c_6 + \delta_3^*\,c_6^2 \qquad \text{for each lens.} \tag{16}$$

This appears like a rather formidable array of equations to solve. But the problem can be tackled by a combination of algebraic and graphical solutions, and enough common sense to realize that there really is little point in trying to find an exact solution for thin lenses anyway. Any solution for thin lenses will be changed as soon as thicknesses are added.

10.3.1.3 The problem is approached by noting that the astigmatism of the (b) lens, C_b, is constant and does not depend on the bending of the lens. This means that Equation (3) in this section can be written in two variables, c_2 and c_6. By using c_2 as a free variable, and choosing a numerical value of c_2, c_6 may be found by solving a quadratic equation. If there are two real solutions one must choose between the positive or negative sign before the square root. In all the work to follow, the positive sign has been taken for the solution. The solutions provided by the negative root are not promising optical systems, because the lens surfaces have too high a curvature.

10.3.1.4 With c_2 and c_6 determined, Equation (2) becomes a linear equation which can be solved for a single value of c_4. Now c_2, c_4, c_6 are determined, so Equations (1) and (4) determine ΣB and ΣE. This procedure may then be repeated for several values of c_2. The values of ΣB and ΣE should then be plotted on a graph with c_2 as the abscissa. A plot of this type is illustrated in Figure 10.4. The ordinates of this graph are $\Sigma B/2\,(n_{k-1}\,u_{k-1}) = {}_3Y_k$ and $\Sigma E/2\Phi$ = F.D., which are the actual transverse third order spherical aberration and the fractional distortion. The thin lens coefficients were computed using $y_1 = 1$ and $\bar{u}_1 = 0.3$. Therefore, the graph in Figure 10.4 shows the spherical aberration and fractional distortion for an $f/5$ system with an image height of 3.0. The graphs in Figure 10.4 show that there are two solutions where the spherical aberration is zero, while the fractional

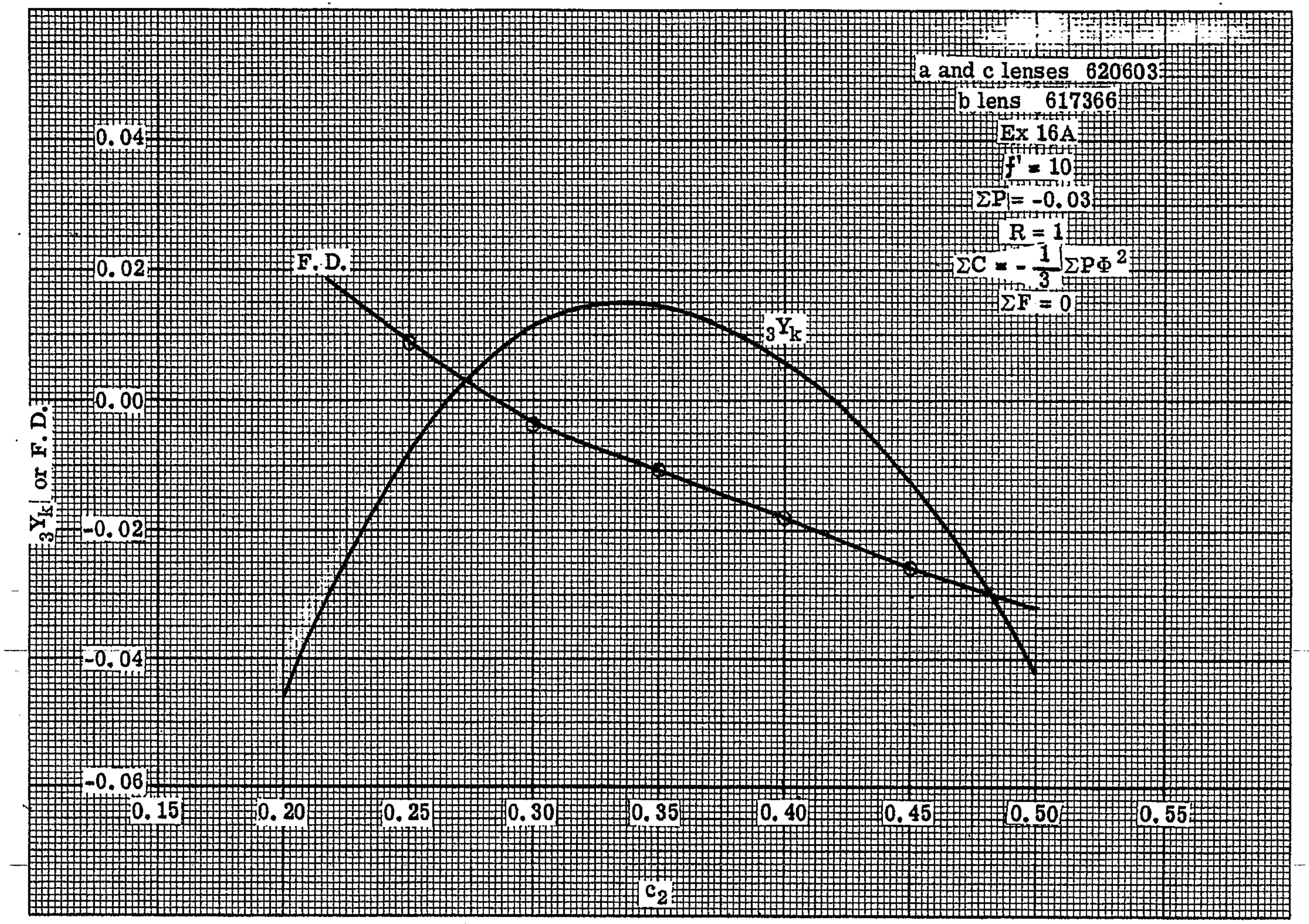


Figure 10.4 - Variation of distortion and spherical aberration with curvature c_2.

distortion is positive for one and negative at the other. The solution with the smaller value of c_2 has the least amount of distortion.

10.3.1.5 Curves of the type shown in Figure 10.4 have been worked out for all the systems included in sheets 1, 2, and 3 of Table 10.6. An attempt to summarize the data is included in Table 10.6, sheets 4 and 5. In nearly every case, the curve for spherical aberration can be represented by a parabola, while the distortion curve can be approximated by a straight line. The data in Table 10.6, sheets 4 and 5, give the information defining the constants of the parabola and the slope of the straight line. For practical purposes, all the parabolas and straight lines can be fitted to the same constants. Therefore,

$$_3Y_k = -2.0\ (c_i - c_{i\ \text{vertex}})^2 + {_3Y_k}_{\ \text{vertex}}\ .$$

In the tables, c_2, c_4, and c_6 are given for the vertex of the parabola. Therefore one can calculate c_2, c_4, and c_6 for any desired $_3Y_k$ from the above equation. Using example 16A in Table 10.6, sheet 5, the values for c_2, c_4, and c_6 for $_3Y_k = 0$ are given by

$$0 = -2.0\ (c_2 - 0.3380)^2 + 0.016, \quad c_2 = 0.249 \text{ and } .427$$

$$0 = -2.0\ (c_4 + 0.2057)^2 + 0.016, \quad c_4 = -0.295 \text{ and } -.116$$

$$0 = -2.0\ (c_6 - 0.0780)^2 + 0.016, \quad c_6 = -0.011 \text{ and } .167$$

10.3.2 Analysis of the data.

10.3.2.1 Notice that the c_2 values for $_3Y_k = 0$, calculated above, do not agree with the data in Figure 10.4, where $c_2 = 0.262$ and 0.416. This is because the equation of the parabola has been simplified to meet all the cases. The slight discrepancy is of little concern at this step of the design, for introducing the thicknesses will change conditions anyway. These figures, therefore, give adequate starting data for the next step of the design. However, before proceeding, the following features of the data in Table 10.6, sheets 4 and 5, should be observed.

(1) Changing the value of R from 1 to 2.0, or from 1 to 0.5, has the effect of moving the parabola downward, with a horizontal vertex shift towards increased c_2 values for $R = 2.0$, and toward decreased c_2 values for $R = 0.5$.

(2) Changing R from 1.0 to 2.0, or from 1 to 0.5, has the effect of moving the F.D. versus c_2 curves upward for $R = 2.0$, and downward for $R = 0.5$ with no appreciable change in slope.

(3) Decreasing ΣP from -0.03 to -0.02 has the effect of moving the parabolas upward, with little effect on the F.D. curves.

10.3.2.2 The $_3Y_k$ and F.D. curves for the same solution for values of $R = 2$ and 0.5, are shown in Figure 10.5. At some value of R (about $R = 0.80$), the distortion curve and the spherical aberration parabola will intersect each other at 0.0 for a c_2 value around 0.27. For an R about 1.5, the curves cross again at 0.0 for a value of $c_2 = 0.35$. This means that if R is variable, there are two solutions corrected for both spherical aberration and distortion. Since ΣP, ΣC, and ΣF are specified for all the curves, these two solutions are then completely corrected to the desired third order aberrations. The solution with the smaller value of c_2 will be referred to as the left hand triplet solution, while the other solution will be called the right hand solution.

10.3.2.3 If glasses with larger $\Delta\nu$ are used, the parabolas are lowered and the two solutions approach each other on the c_2 plot, the final single solutions tend towards a value of R slightly greater than 1.0. The indices of the elements seem to have only a secondary effect on the design while the $\Delta\nu$ difference has a very significant effect.

10.3.3 Ray trace analysis. The designer cannot be sure from the thin lens data how to choose from all the possible choices of glass. There are a very large number of triplets for which the third order distortion and spherical aberration are zero; and the number, of course, is unlimited if distortion residuals are allowed. The only way to really check on the advantage of one design over another is to ray trace the various possibilities. One instinctively feels, however, that if the left hand and right hand solutions can be made to come together that this design will be a good solution. Under this condition the spherical aberration para-

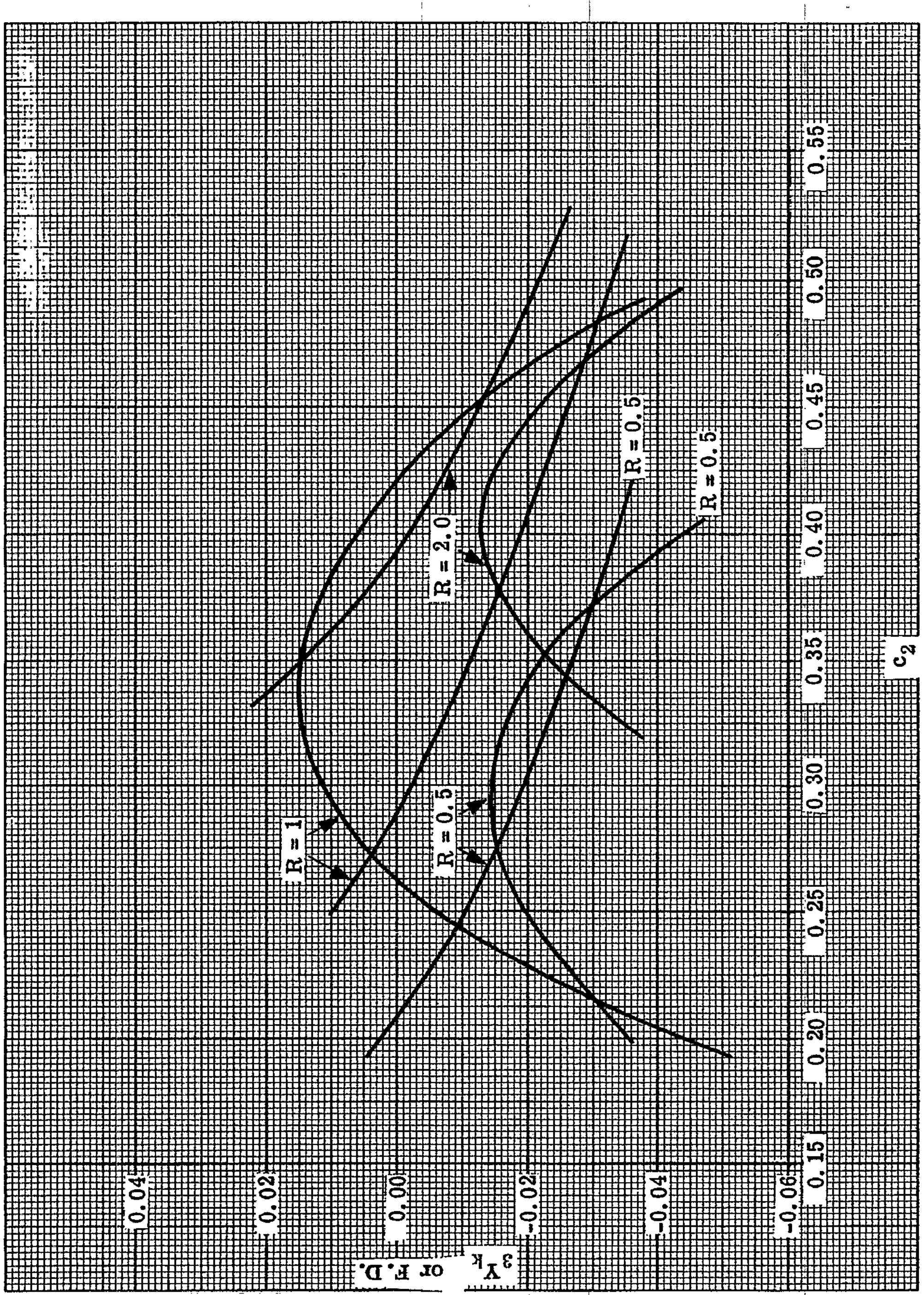


Figure 10.5 - Same plot as in Figure 10.4 with different values of R.

bola and the distortion curve intersect at the maximum of the parabola. Therefore the solution will be somewhat insensitive to changes in curvature, and consequently should be less sensitive to errors in manufacture. Later work on ray trace results will also substantiate that these solutions are preferred to others.

10.3.4 Summary of thin lens design for a triplet objective.

(1) There are two main solutions for $B = F = E = 0$ and $C^* = -\frac{1}{3} P \Phi^2$. One solution is called the left hand solution, $(R < 1)$. The other is the right hand solution, $(R > 1)$.

(2) These solutions are brought together to form a single solution, by increasing the $\Delta\nu$ between the positive elements and the negative element. The R values also converge to a value slightly greater than 1. It is believed that this provides a near optimum solution.

(3) If the $\Delta\nu$ is made too large, there is no solution.

(4) If the value of ΣP is made more positive, the two solutions (if they exist) separate, and in order to bring them together again, the $\Delta\nu$ must be made greater.

10.4 STEP FOUR - THE THICK LENS FIRST ORDER AND THIRD ORDER ABERRATIONS

10.4.1 Lens thickness.

10.4.1.1 Introducing the proper thicknesses in a lens system is also a problem. One has to be sure the positive lenses are of large enough aperture to pass the necessary rays, and the negative lenses have to be made thick enough to resist warping during manufacture. The thickness of the negative lenses can be usually assigned quite easily by adopting the rule that a negative lens should not be thinner than 1/10 its diameter. This usually provides a lens with sufficient strength. In special cases thinner lenses can be made if there is a real need for it; hence this rule is merely a guide.

10.4.1.2 The positive lens thickness is more difficult to ascertain because it depends on the system. It is necessary that the system be almost completely designed before deciding on the diameters of the positive lenses. The designer usually vignettes the oblique beams by cutting the clear aperture of some of the positive lenses. He seldom makes the positive elements with clear apertures large enough to pass the complete oblique beams. Drawings of the lens with pictures of the rays passing through it are very useful in visualizing the thickness required. After the clear apertures of a lens are determined it is still necessary for the diameter to be somewhat larger in order to take care of the edge thickness and mounting rims. When the maximum diameter is known, the thickness is calculated using the thin lens curvatures.

10.4.1.3 Rules for the increased diameter needed to mount the lens vary from shop to shop; thus the problem of lens mounting is a subject in itself and will not be treated here. A designer will have to learn these things through experience, although a shop practice manual may help. The designer must remember, however, that shop people have a natural tendency to resist doing things differently. A designer can miss some good designs if he lets shop people talk him out of a very thin lens, or a glass that is difficult to handle. Formerly designers also had a tendency to resist change, insisting on sticking with their design simply because it was so difficult to recalculate. Today, with modern computing machines, there is no excuse for this. It is now very inexpensive to redesign a system completely just to provide a bit more thickness if it is required by the shop people. There are, however, times when the designer needs a thin element for the reduction of weight or to fit into a tight spot, or an expensive hard-to-handle glass may be required to optimize the design. The designer today can back up his design with proof, so he should be able to violate some shop rules.

10.4.1.4 In order to proceed with the thick lens system, the problem of assignment of thickness will not be discussed further. For the present example thicknesses will be inserted without further explanation.

10.4.1.5 The thin lens first order study of the triplet was started using the glasses 620603 for the crowns, and 617366 for the flint, System No. 16. The thin lens third order study shows that the spherical aberration parabola for these glasses extends far above the zero axis. The two solutions are therefore widely separated. For this reason, it appears that the parabola should probably be lowered. This can be done by choosing a flint with a lower ν value. The thick lens set-up is, therefore, System No. 17 in Table 10.6, sheet 2, with 621362 glass for the negative lens. Using the data from Table 10.6, sheet 5, and the parabolic equation, it is possible to compute the first curvatures for the two solutions with $R = 1$ for zero spherical aberration. These thin lens solutions are as follows:

Left Hand Solution	Right Hand Solution
$c_2 = 0.26$	$c_2 = 0.41$
$c_4 = -0.28$	$c_4 = -0.13$
$c_6 = 0.016$	$c_6 = 0.22$
$t_3 = 1.28$	$t_3 = 1.28$
$t_5 = 1.28$	$t_5 = 1.28$

10.4.1.6 The angle the axial ray makes with the optical axis may be computed from the data in Table 10.6, sheet 2, by setting up a table as in Table 10.4. The values of ϕ_a and t_3 are known; therefore the table is completely determined.

10.4.1.7 As soon as it is necessary to assign thicknesses, the designer has to decide on the f-number and the focal length of the lens. For the following study, it will therefore be assumed that the diameter of the entrance pupil will be 3 and the focal length 10. It is also important to assign a maximum field of obliquity for the lens. Let this (object field) be 20° half angle.

10.4.1.8 The axial paraxial ray should therefore be traced through the system as follows:

$$y_1 = 1.5$$

$$u_o = 0$$

The thin lens axial ray trace for this example then appears as in Table 10.7.

SURFACE	1	2,3	4,5	6,7
$-\phi$	0	−0.184	0.347	−0.210
t		1.24	1.24	
y	1.5	1.5	1.158	1.314
u	0	−0.276	0.126	−0.15

Table 10.7 – Thin lens solution for example 17A in Table 10.6.

10.4.2 Computing the thick lens solution.

10.4.2.1 The thick lens is then set up using the values c_2, c_4, c_6, t_3 and t_5 for either the right hand or the left hand solution with the thicknesses of the lenses inserted. For this example the positive lenses are assigned thicknesses of 0.6 and the negative lens a thickness of 0.25.

10.4.2.2 The second curvatures of the lenses are then computed to maintain the paraxial angles, shown in Table 10.7, between the lenses. The spaces between the lenses may at this time be set at about 1.0.

10.4.2.3 With this initial system, the first order and third order contributions are calculated as shown in Table 8.2.

10.4.3 Iterative analysis and adjustment.

10.4.3.1 As the formalized step-by-step procedure of Section 9 is followed through the remaining steps, it is necessary to examine results and repeat, with changes, earlier steps in order to balance the higher order aberrations. This iterative appraisal and recomputation is the means by which the design can be refined and developed to the desired degree. The mechanics of computation are well described in the foregoing sections and will not be repeated in detail.

10.4.3.2 This discussion, then, will be devoted only to the examination and interpretation of results and to the analytical processes which dictate the iterative changes as design refinement proceeds. With this orientation in mind, and also remembering that by properly programming an automatic computer, the results of one run will produce much of the data necessary for analysis, the discussion will proceed.

10.4.3.3 Next comes the problem of ray tracing, analyzing, and readjusting the lens to the desired third order aberrations. The recommended procedure for doing this is to make small changes in the system and solve Equations 9-(1) through 9-(7). A procedure of this type has been programmed for the I. B. M. 650 computer, and with this program the foregoing triplet system has been studied extensively. A brief account of this study is presented below.

10.4.3.4 First it was decided that the quantities c_2, t_3, c_4, t_5, c_6 would be used as variables. This provided only five variables so it was necessary to provide another variable in order to correct the six quantities B, F, C, P, a, and b. c_5 was used as the extra variable, meaning that the solution departed from $R = 1$. It was possible with three iterations to find the left and right hand solutions, but neither of these solutions were ray traced because it was not possible to tell what value of R the final solution would have. Therefore, it was decided to let c_3 also vary. This provided an extra degree of freedom so the distortion was corrected. In other words, R was allowed to be a variable for the purposes of correcting distortion. In other words, with the seven variables c_2, c_3, t_3, c_4, t_5, c_6, and R, the seven aberration coefficients, B, F, C, P, E, a, and b could be specified. As one would predict from the graphs in Figure 10.5, two solutions were found. This same technique was used in the further study of this lens; hence all the solutions are corrected to exactly zero third order distortion. It was thus possible to compare several designs by varying single parameters, and the third order aberrations could be brought to precisely the required values.

10.5 STEP FIVE - TRACING A FEW SELECTED RAYS

10.5.1 Analysis of the ray tracing results. One finds immediately upon ray tracing that the first and third order aberrations should not be set to zero. The reason for this is that high order aberrations are always present and the third order aberrations have to be set to compensate for them. For example, if the triplet is corrected with $\Sigma B = 0$, the rays traced at $Y = 1.5$ will strike the paraxial image plane at large positive values indicating that the high order aberrations have over-corrected the lens. This is also the reason why ΣP was made equal to -0.03 instead of zero. The same is true with respect to color aberrations. In the triplet it turns out that ΣF, ΣE, and Σb can be set at zero, but the remaining ones, ΣB, ΣC, ΣP, and a have to be set at negative values.

10.5.2 Target values and solutions.

10.5.2.1 Early in the study of this system it was found that the value for ΣP had to be changed from -0.03 to = -0.035 and the spherical aberration had to be under-corrected to $\Sigma B = -0.006$. The first solutions showing interest were computed with the following target values for the third order coefficients:

$$\Sigma B = -0.006$$

$$\Sigma F = 0$$

$$\Sigma C = -\frac{1}{3} P \Phi^2$$

$$\Sigma E = 0$$

$$\Sigma P = -0.035$$

$$\Sigma a = -0.0004$$

$$\Sigma b = 0$$

The chromatic aberrations were left small and unchanged throughout, since the study was done primarily to show how the monochromatic aberrations are corrected.

10.5.2.2 With these target values, two solutions were found. The lens data are included in Table 10.8. The data include the overall thickness T of the lens. Aberration plots similar to Figure 9.1 are shown in Figures 10.6 and 10.7 for the two solutions.

Left Hand Solution		Right Hand Solution	
c	t	c	t
0.2209		0.3116	
	0.600		0.600
-0.0124		-0.0282	
	1.3949		0.8572
-0.2407		-0.1734	
	0.2500		0.2500
0.2606		0.3319	
	0.9631		1.628
0.0777		0.0683	
	0.600		0.600
-0.2822		-0.1891	
	8.453		7.599
R = 0.7685 T = 3.808		R = 1.82 T = 3.935	

Table 10.8 - Left and right hand solutions for a triplet with $\Sigma P = -0.035$.

10.6 STEP SIX - READJUSTING THE THIRD ORDER ABERRATIONS

10.6.1 Analysis of the aberration curves.

10.6.1.1 The aberration curves show that the meridional rays depart from third order drastically at 20°. In the lens designer's language, the tangential field has pulled in rapidly. On the other hand, the sagittal field has moved back relative to the third order field by a much smaller amount. These lenses cannot perform well beyond a 15° half angle. The sagittal rays tend to follow the third order curve much more closely than the meridional rays. Notice also how the skew fan appears similar to the axial fan, but over-corrects in spherical aberration as the field angle increases. The left hand solution appears to be somewhat better than the right hand solution at 20°, but there is little to choose between them at the smaller field angles.

10.6.1.2 It can be seen from Table 10.8 that the left and right hand solutions are widely separated on the c_2 scale. This means the spherical aberration parabola should be lowered. Now this can be done by many methods. One way is to increase $\Delta\nu$ by changing the glass in the flint element. The 649338 glass could be used. However, as the thin lens data indicate (System No. 18), this would lengthen the system and the result would be that the tangential field pulls in even faster. An additional disadvantage is that the 649338 glass lowers the parabola so far that there is no solution with the present glass thicknesses. A second method of lowering the parabola is to make the Petzval sum more negative. One might think at first that this would make the system longer, which is likely to make the tangential field pull in even more, but if the parabola is lowered, the R values will be closer to 1, and the thin lens study showed that for this value of R, the systems are the shortest. Therefore, making $\Sigma P = -0.040$ probably will not make the system much longer.

10.6.2 Readjustment procedure.

10.6.2.1 The value of ΣP was therefore changed to -0.040, and solutions were found with all the other aberrations identically the same. The two solutions and the aberration plots are shown in Table 10.9, and Figures 10.8 and 10.9.

10.6.2.2 The aberration curves now show real improvement. The skew ray fans for the two solutions are quite similar. However the left hand solution appears more symmetrical than the right hand solution, and at 20° it is definitely superior. The R values are now closer together and the barrel length, T, is actually shortened. Notice that the left hand solution is shorter than the right. This may be the reason why the tangential field pulls in further with the right hand solution. Notice also that the left hand solution has a value of R closer to 1.0 than the right. If the parabola were lowered still further, the two solutions would converge to a single solution with $R > 1$ as predicted from the thin lens system.

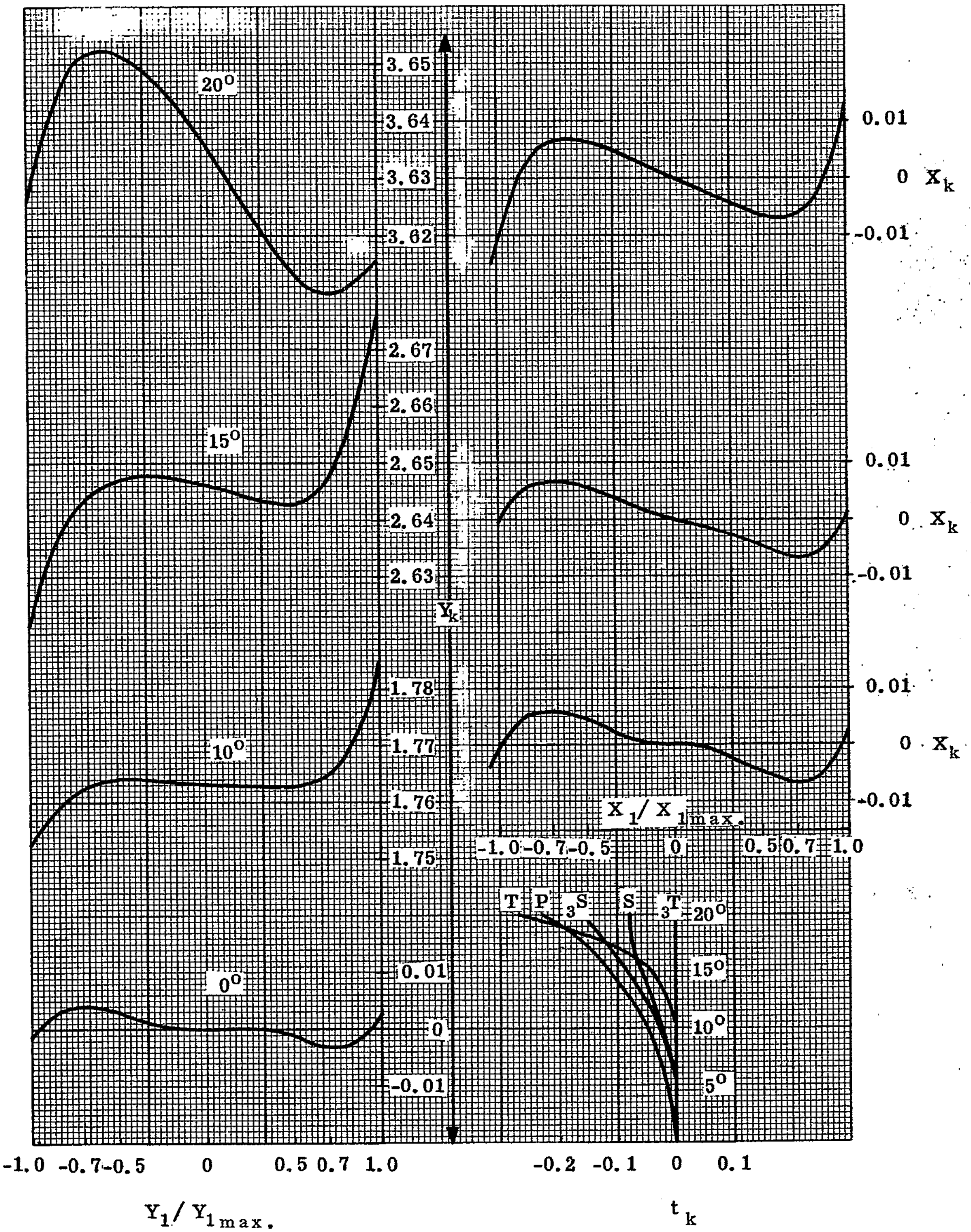


Figure 10.6 - Aberration plots for left hand solution of lenses in Table 10.8.

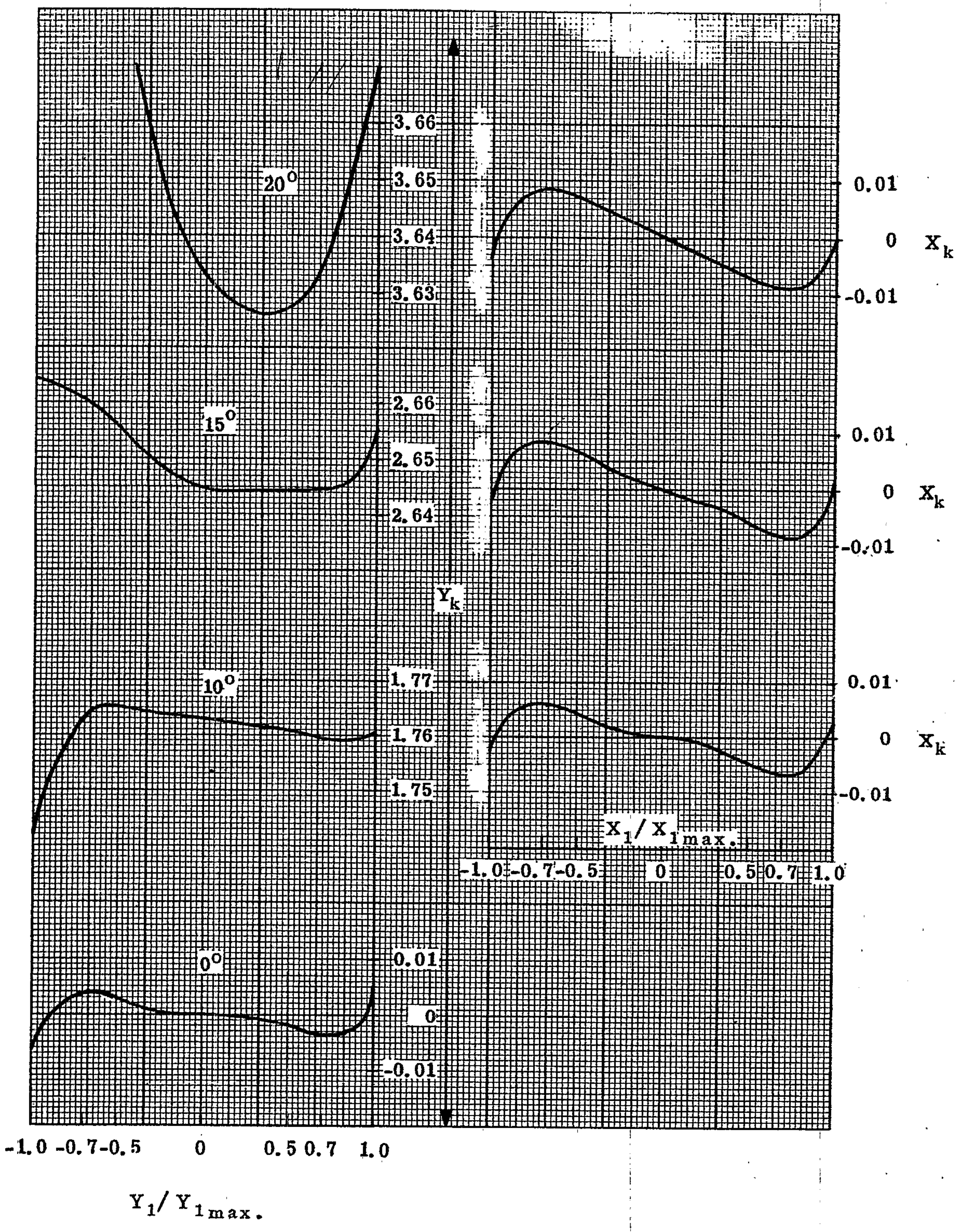


Figure 10.7- Aberration plots for right hand solution of lenses in Table 10.8.

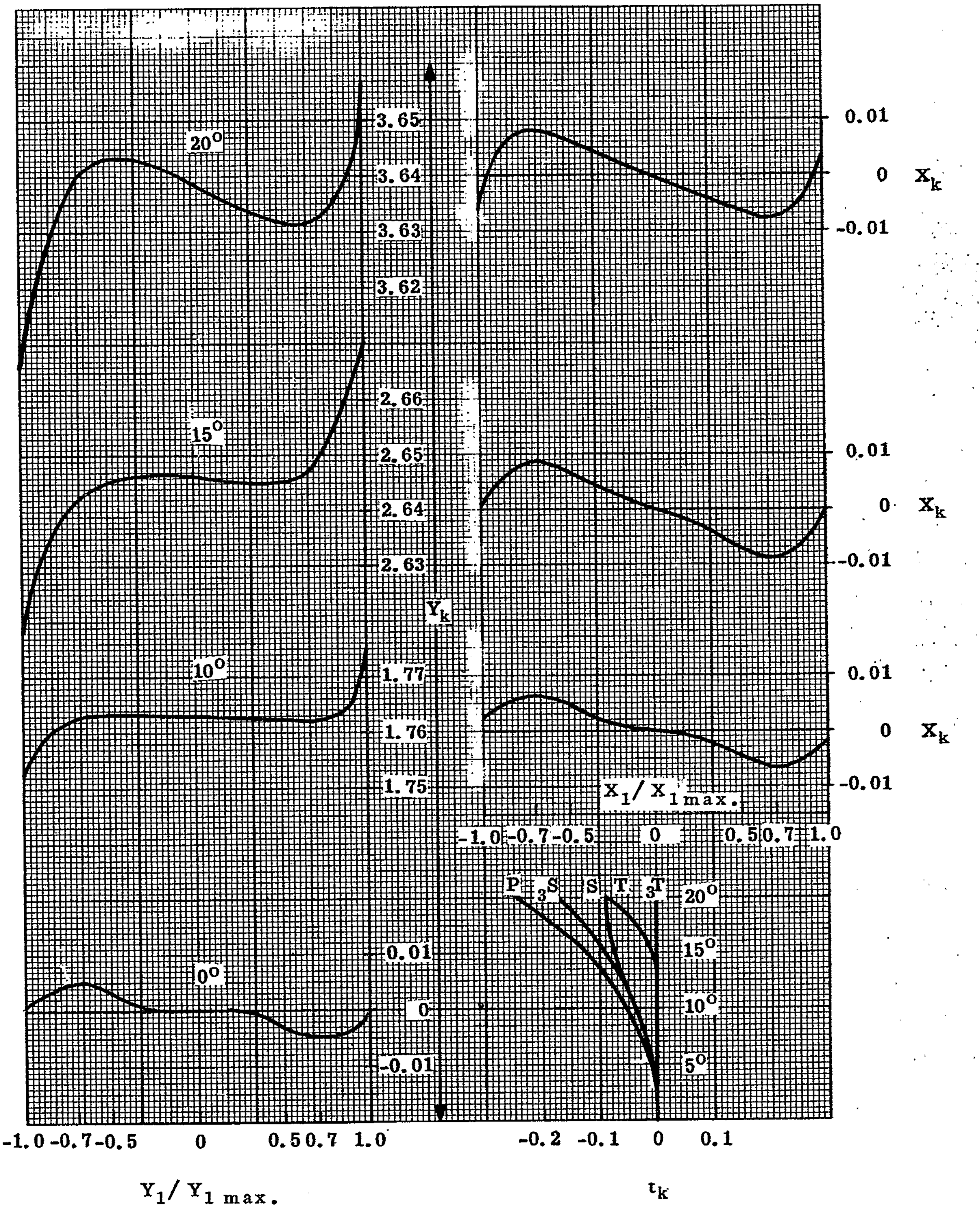


Figure 10.8- Aberration plots for left hand solution of lenses in Table 10.9.

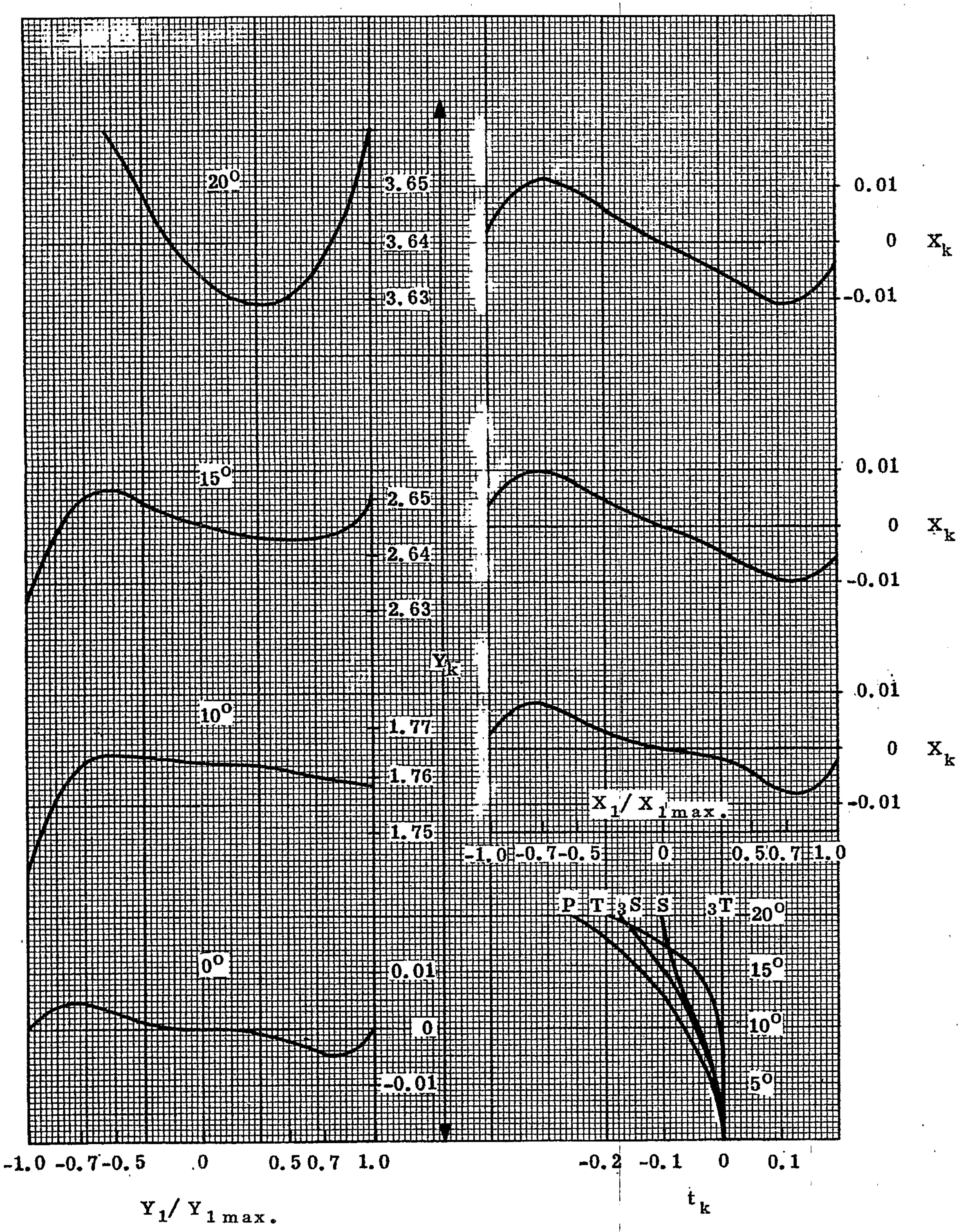


Figure 10.9- Aberration plots for right hand solution of lenses in Table 10.9.

Left Hand Solution c	Left Hand Solution t	Right Hand Solution c	Right Hand Solution t
0.2339		0.2882	
	0.60		0.60
-0.0098		-0.0193	
	1.211		0.8900
-0.2107		-0.1641	
	0.25		0.25
0.2516		0.2998	
	1.016		1.426
0.06799		0.0650	
	0.60		0.60
-0.2556		-0.1963	
	8.363		7.847
R = 0.9003		R = 1.493	
T = 3.677		T = 3.766	

Table 10.9 - Left and right hand solutions for $\Sigma P = -0.040$.

10.6.2.3 At this point, it was noticed that the system still was not as good as that described in Table 8.2. It was finally apparent that the thickness of the negative lens was 0.15, instead of 0.25. This indicated that the aberration parabola was lowered in this solution because of the decreased thickness of the negative lens. Then a new solution was found and the only change was to make $t_b = 0.15$. The result is that left and right hand solutions have R values 0.987 and 1.363. They are drawing closer together and the tangential field does not pull in as rapidly. The surface data for these solutions are included in Table 10.10 (step seven). Figures 10.10 and 10.11 show spot diagrams for these two solutions. In these diagrams only half of the symmetrical image is shown. The diagrams include the appearance of the images as the focal plane is shifted, clearly showing how a shift towards the lens provides a better concentration of light than in the paraxial focus. These diagrams show that there are only slight differences between the imagery in the left and right solutions, out to a half field angle of 15°. However, beyond 15° the left hand solution definitely is superior to that of the right hand. Notice how it shows better concentration and is more symmetrical.

Left Hand Solution c	Left Hand Solution t	Right Hand Solution c	Right Hand Solution t
0.2469		0.283	
	0.60		0.60
-0.00775		-0.01227	
	1.128		0.9289
-0.2024		-0.1692	
	0.15		0.15
0.2568		0.2911	
	1.0738		1.3209
0.0608		0.05869	
	0.60		0.60
-0.2487		-0.2113	
	8.346		8.033
R = 0.987		R = 1.363	
T = 3.552		T = 3.5998	

Table 10.10 - Left and right hand solution for $\Sigma P = -0.040$.

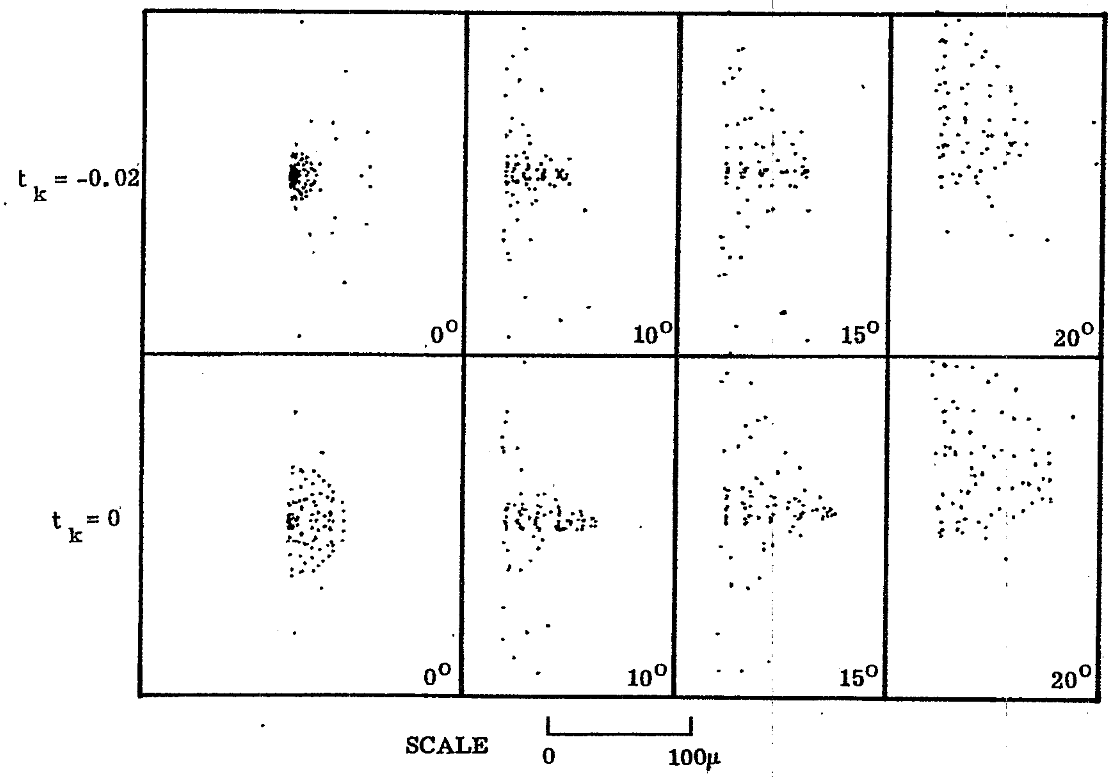


Figure 10.10 - Spot diagrams for left hand solution in Table 10.10.

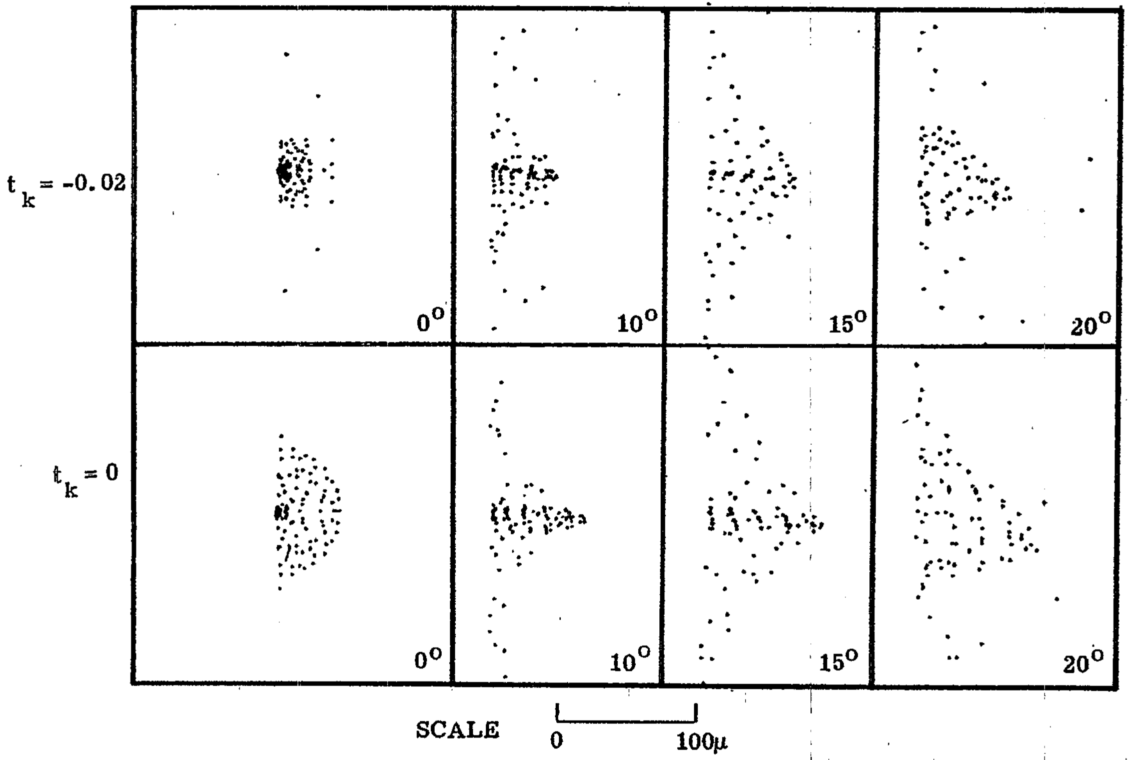


Figure 10.11 - Spot diagrams for right hand solution in Table 10.10.

10.6.3 Analysis of the readjustment procedure.

10.6.3.1 This study showed that reducing the thickness of the negative lens lowers the spherical aberration parabola. This effect then suggested investigating the effect of varying the thicknesses of the positive lenses. The effect is the same, namely lowering the parabola, but not as marked as in the case of the negative lens. This technique of changing the thicknesses of the lenses can be a useful way to compensate for the fact that the parabola is not quite where it should be. The parabola could be lowered still further by reducing the thicknesses, for the case of $\Sigma P = -0.040$, but this is not practical.

10.6.3.2 It is interesting to notice, that in these solutions for the triplet, c_2 is close in value to c_5, and c_4 is close to c_7. For the left hand solution these four surfaces have approximately equal curvatures. It would be very interesting if a solution could be found where all these four curves are identical.

10.7 EVALUATION OF OVER-ALL PERFORMANCE

The design of the optimum triplet is still far from complete, for one must investigate these images carefully by calculating the spot diagrams and energy distributions to be sure the best values for ΣP, ΣC, ΣF and ΣB have been chosen. To do this in detail is an enormous task which realistically can only be done on a very large computer. However with patience and judgment it is possible for designers to arrive at very good solutions.

10.8 SUMMARY

10.8.1 Guide lines. This study has indicated a few guide lines to follow in designing a triplet.

(1) One should always try to design as short a lens as possible to cover a given field.

(2) The spherical aberration parabola can be raised or lowered by the choice of ΣP, thicknesses of lenses, or glass.

(3) It appears that near optimum solutions occur with an R value slightly greater than 1.

10.8.2 Unsolved problems. The study made on this lens merely initiates the reader to the possibilities which need further clarification. A few of the problems are:

(1) What happens as the index of the crown element is increased?

(2) What kinds of solutions can be obtained by first lowering the parabola by glass choice so that there is no solution and then raising the parabola by thickness choice?

(3) What effect has raising the index of the negative lens if $\Delta \nu$ is maintained constant?

(4) When the parabola is too high, is it beneficial to lower it by using aspheric surfaces?

(5) What happens if the number of elements is increased beyond three, and at the same time two or more are cemented together? What happens if the negative element is cemented to either or both positive elements?

(6) If the lens is to cover a wider field, how does one choose the glass?

(7) At finite conjugate will the effect of glass choice and lens thicknesses be the same as at infinite conjugate?

11 TELESCOPE OBJECTIVES

11.1 INTRODUCTION

11.1.1 Scope. Sections 11 through 15 will be devoted to the demonstration of design principles using the simple telescope as a working example. Section 11 will discuss the detailed design of objectives; Section 12, lens erecting systems; Section 13, mirror and prism systems; Section 14, eyepieces; Section 15, complete telescope. It is assumed at this point that the reader has studied all the foregoing material and now has considerable knowledge of optical design. The inexperienced reader should not be concerned if he does not fully grasp the following material on first reading. As he gains experience it will become more and more useful.

11.1.2 The complete telescope. In describing the telescope, it will be assumed that the object is at infinity and that the eyepiece will be focused to place the virtual image at infinity. This means that the distant object will be brought to focus or imaged at the second focal point of the objective and that the first focal point of the eyepiece will coincide with this image. The telescope is then said to be in afocal adjustment.

11.1.3 The Petzval curvature of the system.

11.1.3.1 To start the design of an optical system, one of the first calculations to be made is an estimate of the amount of field curvature in the system, i.e., an estimate of ΣP. This is done by calculating P for each surface using Equation 8-(14), and summing. From Equations 8-(15) and 8-(16), if ΣB, ΣF, and ΣC are all zero, then

$$\alpha_y = \left(\frac{L_{k-1}}{M_{k-1}} - \frac{\bar{L}_{k-1}}{\bar{M}_{k-1}}\right) = \frac{\Sigma P \Phi^2}{2n_{k-1}\ y_{k-1}} \left(\frac{Y_1}{y_1}\right)\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2 {}^* \qquad (1)$$

$$\alpha_x = \frac{K_{k-1}}{M_{k-1}} = \frac{\Sigma P\ \Phi^2}{2n_{k-1}\ y_{k-1}} \left(\frac{X_1}{y_1}\right)\left(\frac{\bar{Y}_o}{\bar{y}_o}\right)^2 . \qquad (2)$$

These equations show that P introduces angular aberrations, α_y and α_x, which are linearly proportional to Y_1/y_1 and to X_1/y_1. This means that for oblique object points the fan of rays entering the objective from a distant object point, do not emerge from the eyepiece as a parallel bundle. Instead the rays come to a finite focus. When $\bar{Y}_o$ is zero, Equations 8-(15) and 8-(16) indicate that the rays from the object point do emerge parallel to the axis because the telescope is in afocal adjustment and $\Sigma B = 0$. For the case described by Equations (1) and (2), the amount of the angular aberration varies as the square of the object height, $\bar{Y}_o$.

11.1.3.2 If the telescope objective is assumed to be thin, then from Equation 8-(28), $P = -\phi/n$. It is possible to change the value of P by using a photographic type of objective, but a simple doublet objective is by far the most common type of lens used in telescopes. In order to derive a general formula for the aberration, it is possible to write $P_o = -\phi_o / \gamma_o$, where γ_o is a constant ranging in value from 1.5 to ∞. For a doublet γ_o is approximately 1.5.

11.1.3.3 The value of P for the eyepiece also depends on how it is designed, but here again it is possible to write $P_e = -\phi_e / \gamma_e$, where ϕ_e is the power of the eyepiece and γ_e is a constant depending on the type of eyepiece. γ_e will range in value from 2.6 to 0.7 for most eyepieces.

11.1.3.4 Using these values of P for the objective and the eyepiece, it is possible to compute, from Equation (1), the angular aberration,

$$\alpha_y = -\frac{n_o^2\ \ u_o^2}{2\,n_{k-1}\ \ y_{k-1}} \left(\bar{Y}_o^2\right)\left(\frac{Y_1}{y_1}\right)\left(\frac{\phi_o}{\gamma_o} + \frac{\phi_e}{\gamma_e}\right) .$$

Since a telescope is usually used in air, n_o and n_{k-1} are 1. Then,

$$\alpha_y = -\frac{1}{2}\left(\frac{\bar{Y}_o}{t_o}\right)^2 \left(\frac{Y_1}{y_1}\right) MP^2\ \ y_{k-1}\ \ \phi_e \left(\frac{-1}{\gamma_o\ MP} + \frac{1}{\gamma_e}\right) . \qquad (3)$$

---

*For a telescope with an object point at infinity one should strictly speaking use $(\tan U_o)/u_o$ instead of $\bar{Y}_o/\bar{y}_o$ but it is satisfactory to use the above equations and assume that the object distance is very large but finite.

This equation shows that the angular aberration due to Petzval curvature only varies

(1) as the square of the object height,

(2) approximately as the square of the MP,

(3) linearly with y_{k-1} .

11.1.3.5 For a given $(\overline{Y}_o / t_o)$ MP, the angular aberration in a telescope can be made small by making ϕ_e or y_{k-1} small. It is customary to specify the angular aberration of telescopes in diopters using the definition $d = 100\alpha / y_{k-1}$, where d is in diopters when y_{k-1} is given in centimeters and α is expressed in radians.

11.1.3.6 As a numerical example consider Table 11.1 which shows the values of α in minutes of arc and in diopters for telescopes of three different magnifications. To make these calculations the following assumptions were made:

$y_{k-1} = -0.35$ cm

$f_e = 2.00$ cm

$(\overline{Y}_o / t_o)\ \mathrm{MP} = -0.6$

$Y_1 / y_1 = 1.0$

$\gamma_o = \gamma_e = 1.5$

The angular aberrations in Table 11.1 are all positive which means that the oblique bundles are focused behind the observer's eye making it impossible for him to focus on the image. It also means that the eyepiece must be focused towards the objective in order to remove the angular aberration for the off-axis object point. If it is moved towards the objective then the telescope will have angular aberration of the opposite sign for the central image. This indicates that the observer can accommodate and completely focus-out the angular error. The large angular aberration due to field curvature in telescopes is therefore not as serious as it may appear for it is possible for the eye to change focus as the observer views different parts of the field.

MP $\frac{\overline{Y}_o}{t_o} \cdot$ MP	-2	-5	-10	
-0.6	-9.0	-7.2	-6.6	Diopters
-0.6	107'	87'	79'	Minutes

Table 11.1- Angular aberration for telescopes of three magnifying powers.

11.1.3.7 In military telescopes it is often necessary to insert a reticle in the focal plane of the objective. Since a reticle is used to measure distances in the object space, it is important to design the objective with a flat field on the reticle, which usually means that the lens has to be more complex than the usual telescope doublet objective. Because the Petzval curvature of the eyepiece cannot be made zero, the eyepiece cannot focus the entire reticle with a single setting. Hence the reticle may not appear perfectly sharp, but if the objective is well corrected there is no parallax between the object and the reticle.

11.1.3.8 If a reticle is not needed in the design there is usually very little need to attempt to reduce the Petzval curvature of the objective by using a compound photographic lens type of objective. Equation (3) shows that the objective contribution, γ_o , is multiplied by the magnifying power of the telescope. For high power telescopes therefore, the objective adds a negligible amount of field curvature. This is why the majority of telescopes use simple doublets for the objective. If the power of the telescope is low then one must consider using some type of lens other than a doublet objective.

11.2 DESIGN PROCEDURE FOR A THIN LENS TELESCOPE OBJECTIVE

11.2.1 First order, thin lens.

11.2.1.1 The doublet, of course, consists of two lenses, and one can immediately start to fill out a table as was done with the triplet objective in Section 10.2. This has been done in Table 11.2.

	Lens a	Lens b	Image Plane
$-\phi$	$-\phi_a$	$-\phi_b$	0
t		0	$\frac{1}{\phi_a + \phi_b}$
y	1	1	0
u	0	$-\phi_a$	$-\phi_a - \phi_b$
$\bar{y}$	0	0	$\frac{1}{\phi_a + \phi_b}$
$\bar{u}$	1	1	1
ν	ν_a	ν_b	0
$-\phi y^2/\nu$	$-(\phi/\nu)_a$	$-(\phi/\nu)_b$	$\Sigma a = -\frac{\phi_a}{\nu_a} - \frac{\phi_b}{\nu_b}$
$\frac{-\phi y \bar{y}}{\nu}$	0	0	$\Sigma b = 0$
$-\phi/n$	$-\phi_a/n_a$	$-\phi_b/n_b$	$\Sigma P = -\frac{\phi_a}{n_a} - \frac{\phi_b}{n_b}$

Table 11.2 - Computing table for the thin lens telescope doublet.

11.2.1.2 In order to solve for ϕ and have the axial color zero, assuming the two elements are close together, the following two equations must be satisfied:

$$\phi_a + \phi_b = \phi , \tag{4}$$

and

$$\frac{\phi_a}{\nu_a} + \frac{\phi_b}{\nu_b} = 0 . \tag{5}$$

The solution of these equations is

$$\phi_a = \phi \frac{\nu_a}{\nu_a - \nu_b} , \tag{6}$$

and

$$\phi_b = \phi \frac{\nu_b}{\nu_b - \nu_a} . \tag{7}$$

By using these equations, the value of ΣP from Equation 8-(28) is

$$\Sigma P = -\phi \left(\frac{\nu_a / n_a - \nu_b / n_b}{\nu_a - \nu_b} \right) \quad . \tag{8}$$

11.2.1.3 These equations show that any two glasses with a difference in ν may be used to design a doublet. As will be seen later however, $\Delta \nu$ should be large in order to reduce the monochromatic aberrations. In principle the ΣP may be made equal to zero by the proper choice of glass. Actually, the ratio

$$\frac{\nu_a / n_a - \nu_b / n_b}{\nu_a - \nu_b}$$

is nearly constant for any glasses chosen with a reasonable value of $(\nu_a - \nu_b)$. It is therefore not practical to attempt to reduce P in a doublet by choosing the proper glasses. Once the two glasses for the doublet are chosen the following is known about the lens:

(1) The focal lengths of the (a) and (b) lenses, using Equations (6) and (7).

(2) The axial color, which was set equal to zero using Equation 6-(42).

(3) The transverse color, which is zero, using Equation 6-(41), because the objective is the entrance pupil.

(4) The Petzval curvature, using Equation (8).

(5) The third order astigmatism, using Equation 8-(26).

(6) The third order distortion, using 8-(27).

Only two aberrations of the third order, B and F, remain uncorrected.

11.2.2 Third order, thin lens.

11.2.2.1 As explained in Section 8.10.1, it is possible to compute the coefficients (α and β) for the following thin lens equations:

$$B_a = \alpha_{1a} + \alpha_{2a} c_1 + \alpha_{3a} c_1^2 \tag{9}$$

$$B_b = \alpha_{1b} + \alpha_{2b} c_3 + \alpha_{3b} c_3^2 \tag{10}$$

$$F_a = \beta_{1a} + \beta_{2a} c_1 \tag{11}$$

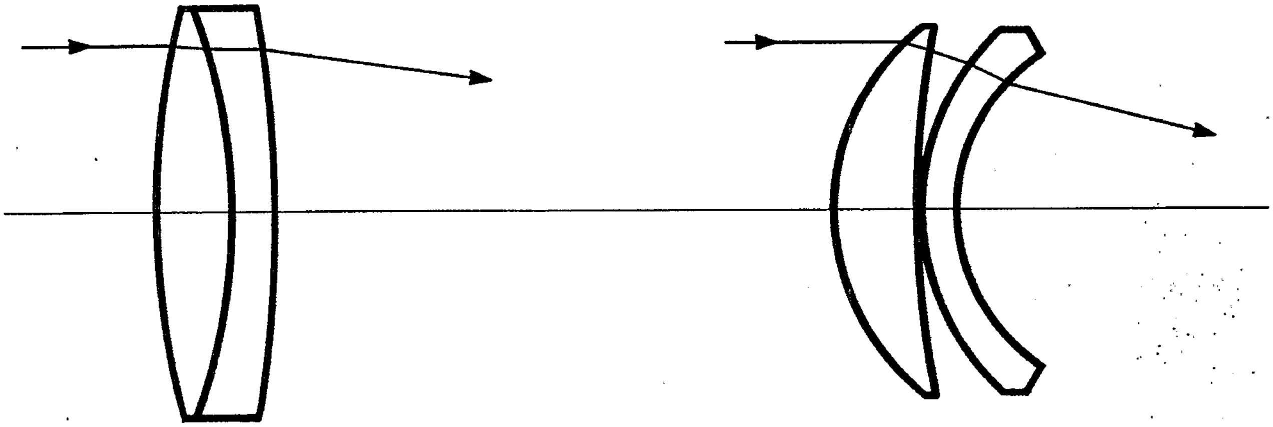
$$F_b = \beta_{1b} + \beta_{2b} c_3 \tag{12}$$

11.2.2.2 By setting $B_a + B_b = 0$ and $F_a + F_b = 0$, the above equations may be reduced to a second degree equation in c_1. There are then two real solutions called the left and right hand solutions. Examples of the two solutions are shown in Figure 11.1. The doublet used in the example was computed for the following glasses:

Lens (a) $n_D = 1.511$ $\nu = 63.5$

(b) $n_D = 1.649$ $\nu = 30.6$

11.2.2.3 The doublets shown in Figure 11.1 have the low dispersion glass in the front element facing the infinite conjugate side of the lens. Doublet solutions can equally well be found with the high dispersion glass in the front. Figure 11.2 shows the left hand solution for the same glasses with the positions reversed. The left hand solutions with the positive lens in front have the most favorable shape for the passage of the axial rays. Therefore most telescope objectives are solutions of this type, and they are referred to as Fraunhofer objectives.



Left hand solution

Right hand solution

Figure 11.1 - Two types of doublets. For both types, the positive lens is in front and is of low dispersion glass.

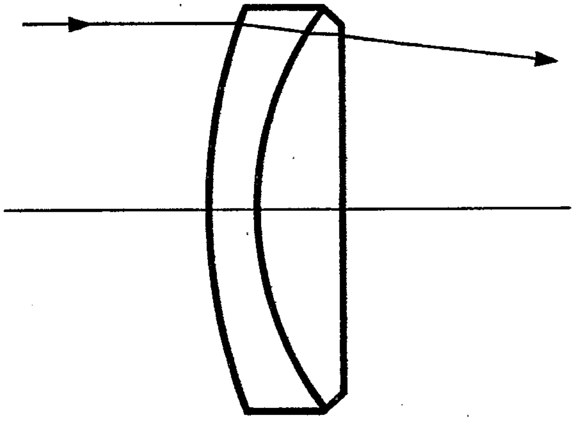


Figure 11.2 - Doublet with negative lens in front. The glasses are the same as the left hand solution in Figure 11.1 with the positions reversed.

11.2.3 Thin lens solutions.

11.2.3.1 Table 11.3 contains a selection of left hand solutions for 38 thin lens systems of the Fraunhofer type. Table 11.4 includes a list of the left hand solution for lenses with the negative lens in front. Systems with the negative lens in front appear to offer no advantages over the Fraunhofer type and so will not be discussed further.

11.2.3.2 There are of course numerous other combinations of glass one can pick, but those shown in Table 11.3 provide a sufficient variety of choices to build up an understanding of the solutions. The following points should be noted about the solutions:

(1) The series of solutions 1 to 8 have the same glass for the positive lens. The negative lenses however are all selected from the ordinary glass line shown in Figure 6.18. Notice that the first curvature for each solution is approximately 0.165 and the last curvature is about -0.08. The outside appearance of these lenses is therefore very similar. The difference between the lenses is in the curvatures of the internal surfaces.

(2) When combinations of glasses with small ν differences are selected, the curvature of the second surface is stronger than that of the third surface. This means that the surfaces do not edge contact and a spacer is needed. As the ν difference is increased, the difference in curvature of these two surfaces decreases, and at large ν differences the third surface becomes stronger than the second. The lenses then contact at the edge.

11.3 DESIGN PROCEDURE FOR A THICK LENS TELESCOPE OBJECTIVE

11.3.1 The thick lens doublet.

11.3.1.1 Very little choice can be made, from the thin lens data alone, between the numerous telescope objectives. In order to make a selection from the possible glass choices, it is necessary to trace rays through the designs.

11.3.1.2 As was pointed out in Section 10.4.1.7, when a lens is studied by ray tracing it is necessary to decide on a definite f - number. Changing the f - number requirement changes completely the conclusions one draws from the ray tracing. In order to illustrate the design procedure, a study will be made for a particular lens of f - number 3.57 with a focal length of 10.

11.3.2 Automatic correction of the third order aberrations.

11.3.2.1 After thicknesses are added to the thin lens solutions, it is necessary to readjust the curvatures to correct the spherical aberration, coma, and axial color to the required residuals. One cannot set the third order aberrations to zero if the total aberrations are to be zero, for the actual ray tracing will reveal the presence of higher order aberration.

11.3.2.2 In order to compare various glass choices in doublets it is necessary to hold constant many variables. In the study to be described the following parameters were held constant.

(a) The focal length of each lens was 10.0.

(b) The marginal ray entered the lens at $Y_1 = 1.4$.

(c) The third order spherical aberration was adjusted until the marginal ray at $Y_1 = 1.4$ in D light focused at the paraxial D focus.

(d) The axial first order color was adjusted until the rays traced at a value of $Y_1 = 1.0$ for F and C light united in the paraxial focal plane.

(e) The thickness of the positive and negative element was made 0.5 and 0.3, respectively

(f) The space between the lenses was maintained automatically. There were two alternatives. Where $r_2 > r_3$, the elements were spaced by 0.01 at the aperture height of $y \doteq 1.4$; where $r_2 < r_3$, a space of 0.01 was set at the vertex.

(g) The third order coma was corrected to zero in all cases.

Glass values Element		Total curvatures		Individual surface curvatures				Case No.
n_{Da} ν_a a	n_{Db} ν_b b	c_a	c_b	c_1	c_2	c_3	c_4	
1.511 63.5	1.5795 41.0	0.5523	−0.3145	0.1652	−0.3871	−0.3802	−0.0658	1
1.511 63.5	1.621 36.2	0.4552	−0.2135	0.1658	−0.2893	−0.2848	−0.0713	2
1.511 63.5	1.649 33.8	0.4184	−0.1754	0.1660	−0.2525	−0.2499	−0.0746	3
1.511 63.5	1.689 30.9	0.3812	−0.1376	0.1659	−0.2152	−0.2163	−0.0787	4
1.511 63.5	1.720 29.3	0.3634	−0.1190	0.1659	−0.1975	−0.2009	−0.0819	5
1.511 63.5	1.8052 25.5	0.3270	−0.0833	0.1656	−0.1614	−0.1730	−0.0896	6
1.511 63.5	1.5704 48.1	0.8070	−0.5476	0.1563	−0.6507	−0.6351	−0.0875	7
1.511 63.5	1.5838 46.0	0.7101	−0.4503	0.1595	−0.5506	−0.5354	−0.0851	8
1.511 63.5	1.605 43.6	0.6245	−0.3622	0.1620	−0.4625	−0.4476	−0.0854	9
1.511 63.5	1.617 36.6	0.4798	−0.2353	0.1657	−0.3141	−0.3077	−0.0723	10
1.511 63.5	1.717 48.2	0.8123	−0.4394	0.1411	−0.6711	−0.6259	−0.1865	11
1.511 63.5	1.720 50.3	0.9415	−0.5293	0.1228	−0.8186	−0.7630	−0.2337	12
1.511 63.5	1.8037 41.8	0.5727	−0.2397	0.1631	−0.4095	−0.3779	−0.1382	13
1.517 64.5	1.689 30.9	0.3713	−0.1335	0.1648	−0.2065	−0.2091	−0.0756	14
1.611 58.8	1.5795 41.0	0.5516	−0.4090	0.1594	−0.3922	−0.4009	0.0080	15
1.611 58.8	1.617 36.6	0.4397	−0.2734	0.1538	−0.2859	−0.2903	−0.0170	16
1.611 58.8	1.621 36.2	0.4318	−0.2638	0.1535	−0.2782	−0.2824	−0.0186	17
1.611 58.8	1.649 33.8	0.3895	−0.2126	0.1525	−0.2369	−0.2403	−0.0277	18
1.611 58.8	1.689 30.9	0.3482	−0.1637	0.1519	−0.1963	−0.2002	−0.0366	19
1.611 58.8	1.720 29.3	0.3290	−0.1403	0.1518	−0.1773	−0.1819	−0.0416	20
1.611 58.8	1.689 30.9	0.3448	−0.1607	0.1519	−0.1930	−0.1973	−0.0366	21
1.620 60.3	1.617 36.6	0.3910	−0.2308	0.1521	−0.2389	−0.2459	−0.0151	22
1.620 60.3	1.649 33.8	0.3526	−0.1827	0.1510	−0.2016	−0.2080	−0.0253	23
1.620 60.3	1.7506 27.8	0.2904	−0.1067	0.1500	−0.1404	−0.1503	−0.0437	24
1.620 60.3	1.8052 25.5	0.2731	−0.0861	0.1498	−0.1232	−0.1362	−0.0502	25
1.620 60.3	1.8037 41.8	0.4902	−0.2537	0.1490	−0.3412	−0.3253	−0.0716	26
1.638 55.5	1.617 36.6	0.4603	−0.3139	0.1527	−0.3075	−0.3144	−0.0005	27
1.638 55.5	1.649 33.8	0.4009	−0.2400	0.1503	−0.2506	−0.2551	−0.0151	28
1.638 55.5	1.689 30.9	0.3536	−0.1823	0.1493	−0.2044	−0.2083	−0.0260	29
1.638 55.5	1.720 29.3	0.3308	−0.1542	0.1490	−0.1818	−0.1861	−0.0319	30
1.5286 51.6	1.5497 45.8	1.6831	−1.437	0.1206	−1.563	−1.548	−0.1113	31
1.5286 51.6	1.5795 41.0	0.9209	−0.6675	0.1481	−0.7729	−0.7568	−0.0893	32
1.5286 51.6	1.621 36.2	0.6339	−0.3785	0.1584	−0.4755	−0.4604	−0.0819	33
1.5286 51.6	1.649 33.8	0.5484	−0.2926	0.1610	−0.3874	−0.3736	−0.0810	34
1.5286 51.6	1.689 30.9	0.4716	−0.2167	0.1629	−0.3087	−0.2975	−0.0808	35
1.5286 51.6	1.72 29.3	0.4378	−0.1825	0.1635	−0.2742	−0.2646	−0.0822	36
1.5286 51.6	1.72 42.0	0.4981	−0.2268	0.1623	−0.3357	−0.3193	−0.0926	37
1.5286 51.6	1.76 29.3	0.4378	−0.1729	0.1639	−0.2738	−0.2620	−0.08913	38

Table 11.3 - Thin lens aplanatic doublets of focal length 10.

Glass values Element		Total curvatures		Individual surfaces curvatures				Case No.
n_{Da}^{a} νa	n_{Db}^{b} νb	c_a	c_b	c_1	c_2	c_3	c_4	
1.5795 41.	1.511 63.5	-0.3145	0.5523	0.2270	0.5415	0.5473	-0.0050	1
1.621 36.2	1.511 63.5	-0.2135	0.4552	0.2332	0.4467	0.4496	-0.0056	2
1.649 33.8	1.511 63.5	-0.1754	0.4184	0.2369	0.4123	0.4127	-0.0057	3
1.689 30.9	1.511 63.5	-0.1376	0.3812	0.2417	0.3793	0.3755	-0.0057	4
1.720 29.3	1.511 63.5	-0.1190	0.3634	0.2453	0.3643	0.3577	-0.0057	5
1.8052 25.5	1.511 63.5	-0.0833	0.3270	0.2543	0.3376	0.3217	-0.0053	6
1.5704 48.1	1.511 63.5	-0.5476	0.8070	0.2487	0.7962	0.8109	0.0039	7
1.5838 46.0	1.511 63.5	-0.4503	0.7101	0.2465	0.6968	0.7108	0.0007	8
1.605 43.6	1.511 63.5	-0.3622	0.6245	0.2471	0.6092	0.6227	-0.0018	9
1.617 36.6	1.511 63.5	-0.2353	0.4798	0.2342	0.4695	0.4744	-0.0055	10
1.717 48.2	1.511 63.5	-0.4394	0.8123	0.3500	0.7895	0.8315	0.0193	11
1.720 50.3	1.511 63.5	-0.5293	0.9415	0.3974	0.9267	0.9791	0.0376	12
1.8037 41.8	1.511 63.5	-0.2397	0.5727	0.3029	0.5426	0.5699	-0.0027	13
1.689 30.9	1.517 64.5	-0.1335	0.3713	0.2385	0.3718	0.3666	-0.0045	14
1.5795 41.0	1.611 58.8	-0.4090	0.5516	0.1532	0.5622	0.5540	0.0023	15
1.617 36.6	1.611 58.8	-0.2734	0.4397	0.1788	0.4521	0.4476	0.0079	16
1.621 36.2	1.611 58.8	-0.2638	0.4318	0.1805	0.4443	0.4399	0.0082	17
1.649 33.8	1.611 58.8	-0.2126	0.3895	0.1900	0.4026	0.3986	0.0092	18
1.689 30.9	1.611 58.8	-0.1637	0.3482	0.1994	0.3631	0.3580	0.0098	19
1.720 29.3	1.611 58.8	-0.1403	0.3290	0.2049	0.3452	0.3390	0.0100	20
1.689 30.9	1.611 58.8	-0.1607	0.3448	0.1995	0.3602	0.3548	0.0098	21
1.617 36.6	1.620 60.3	-0.2308	0.3910	0.1769	0.4077	0.4007	0.0097	22
1.649 33.8	1.620 60.3	-0.1827	0.3526	0.1876	0.3703	0.3634	0.0108	23
1.7506 27.8	1.620 60.3	-0.1067	0.2904	0.2074	0.3141	0.3023	0.0118	24
1.8502 25.5	1.620 60.3	-0.0861	0.2731	0.2147	0.3007	0.2851	0.0121	25
1.8037 41.8	1.620 60.3	-0.2537	0.4902	0.2361	0.4898	0.5031	0.0130	26
1.617 36.6	1.638 55.5	-0.3139	0.4603	0.1623	0.4762	0.4696	0.0094	27
1.649 33.8	1.638 55.5	-0.2400	0.4009	0.1774	0.4174	0.4127	0.0118	28
1.689 30.9	1.638 55.5	-0.1823	0.3536	0.1889	0.3712	0.3665	0.0128	29
1.720 29.3	1.638 55.5	-0.1542	0.3308	0.1951	0.3494	0.3439	0.0131	30
1.5497 45.8	1.5286 51.6	-1.437	1.6831	0.2721	1.7087	1.7231	0.0399	31
1.5795 41.0	1.5286 51.6	-0.6675	0.9209	0.2505	0.9180	0.9333	0.0124	32
1.621 36.2	1.5286 51.6	-0.3785	0.6339	0.2438	0.6223	0.6360	0.0021	33
1.649 33.8	1.5286 51.6	-0.2926	0.5484	0.2433	0.5359	0.54795	-0.00047	34
1.689 30.9	1.5286 51.6	-0.2167	0.4716	0.2438	0.4604	0.4692	-0.0024	35
1.72 29.3	1.5286 51.6	-0.1825	0.4378	0.2456	0.4281	0.4347	-0.0030	36
1.72 42.0	1.5286 51.6	-0.2268	0.4981	0.2560	0.4827	0.4962	-0.0018	37
1.76 29.3	1.5286 51.6	-0.1729	0.4378	0.2531	0.4260	0.4344	-0.0034	38

Table 11.4 - Telescope objectives with flint in front.

11.3.2.3 In designing a doublet of a particular glass choice, it is necessary to estimate the values at which to set the third order spherical aberration, B, and axial color, a. Then surfaces 1, 2, and 3 are varied to provide derivatives for B, F, and a. The following three equations are then solved for the required values of B, F, and a.

$$\Delta \Sigma B = \frac{\partial \Sigma B}{\partial c_1} \Delta c_1 + \frac{\partial \Sigma B}{c_2} \Delta c_2 + \frac{\partial \Sigma B}{\partial c_3} \Delta c_3 ,$$

$$\Delta \Sigma F = \frac{\partial \Sigma F}{\partial c_1} \Delta c_1 + \frac{\partial \Sigma F}{\partial c_2} \Delta c_2 + \frac{\partial \Sigma B}{\partial c_3} \Delta c_3 ,$$

$$\Delta \Sigma a = \frac{\partial \Sigma a}{\partial c_1} \Delta c_1 + \frac{\partial \Sigma a}{\partial c_2} \Delta c_2 + \frac{\partial \Sigma a}{\partial c_3} \Delta c_3 .$$

This is the method described in Section 9.2.4.12. Since the changes are not linear, it is necessary to repeat the procedure for several iterations. When the desired third order values are found, the solution is then ray traced in D, F, and C light at 0° with values of $Y_1 = 1.4$, 1.2, 1.0, and 0.8. From this ray tracing data it is possible to determine if condition c and d are fulfilled. If they are not it is necessary to assign a new value to B and a and repeat the process.

11.3.2.4 Tables 11.5, 11.6, and 11.7 show the data for three doublets designed in this manner. The design shown in Table 11.7 illustrates a careful balance of high order aberrations for D light. In order to arrive at this design it was necessary to choose just the right ν number for the negative lens. If a flint with a larger ν number had been chosen, the rays at an aperture of $Y_1 = 1.4$ would have crossed the paraxial focal plane at a larger negative value and this would have been impossible to correct without introducing a large positive zonal aberration. Notice how the F light starts out to be under-corrected (negative Y_k), but as Y_1 is increased it becomes over-corrected (positive Y_k). The C light starts out positive and then turns toward the negative side. This is evidence of chromatic variation of spherical aberration. Note also how the aberration for the lenses in Tables 11.5 and 11.6 are larger than in Table 11.7 even though all the curvatures are smaller. If a different f - number is needed one would choose a different flint element for optimum correction.

11.3.2.5 The aberration curves in Figure 11.3 are for the ray data given in Table 11.6. The D light curve is typical for a telescope objective. The third order spherical aberration is undercorrected. The curve, for small values of Y_1, starts out below the reference axis following the third order aberration, but it then starts to depart and swings towards the positive side. This is due to higher order aberrations which, in this case, are positive. By evaluating the constants in Equation 8-(1) for this aberration curve we find that

$$b_o = 0, \quad b_3 = -0.0014, \quad \text{and} \quad b_5 = 0.000619$$

Lens specifications				
c	t	n	ν	
0.1672				
	0.5	1.511	63.5	$f' = 9.995478$
-0.1594				
	0.0218	1.0		$l' = 9.624807$
-0.1709				
	0.3	1.80489	25.4	
-0.0886				
Ray-trace data				
Y_1	Y_D	Y_F	Y_C	
1.4	0.000279	0.002047	0.000618	Y is the height of the ray in the D light paraxial focal plane.
1.2	-0.000689	0.000362	-0.000233	
1.0	-0.000777	-0.000202	-0.000291	
0.8	-0.000542	-0.000267	-0.000089	

Table 11.5 - Lens specification and ray trace data for an achromatic doublet with large ν difference.

Lens specifications				
c	t	n	ν	
		1		
0.167876				
	0.5	1.511	63.5	$f' = 10.000000$
-0.244936				
	0.01	1.0		$l' = 9.597161$
-0.243789				
	0.3	1.649	33.8	
-0.073862				
Ray-trace data				
Y_1	Y_D	Y_F	Y_C	
1.4	0.000201	0.002359	0.000334	
1.2	-0.000423	0.000786	-0.000079	
1.0	-0.000526	0.000060	-0.000089	
0.8	-0.000387	-0.000176	0.000055	

Table 11.6 - Lens specification and ray trace data for an achromatic doublet with moderate ν difference.

Lens specifications				
c	t	n	ν	
0.168413				
	0.5	1.511	63.5	$f' = 10.0000$
-0.290972				
	0.01	1.0		$l' = 9.578138$
-0.289068				
	0.3	1.605	38.0	
-0.067287				
Ray-trace data				
Y_1	Y_D	Y_F	Y_C	
1.4	-0.000366	0.001065	0.000038	
1.2	+0.000083	0.000618	0.000673	
1.0	-0.000034	-0.000064	0.000628	
0.8	-0.000092	-0.000411	0.000542	

Table 11.7 - Lens specification and ray trace data for an achromatic doublet with small ν difference.

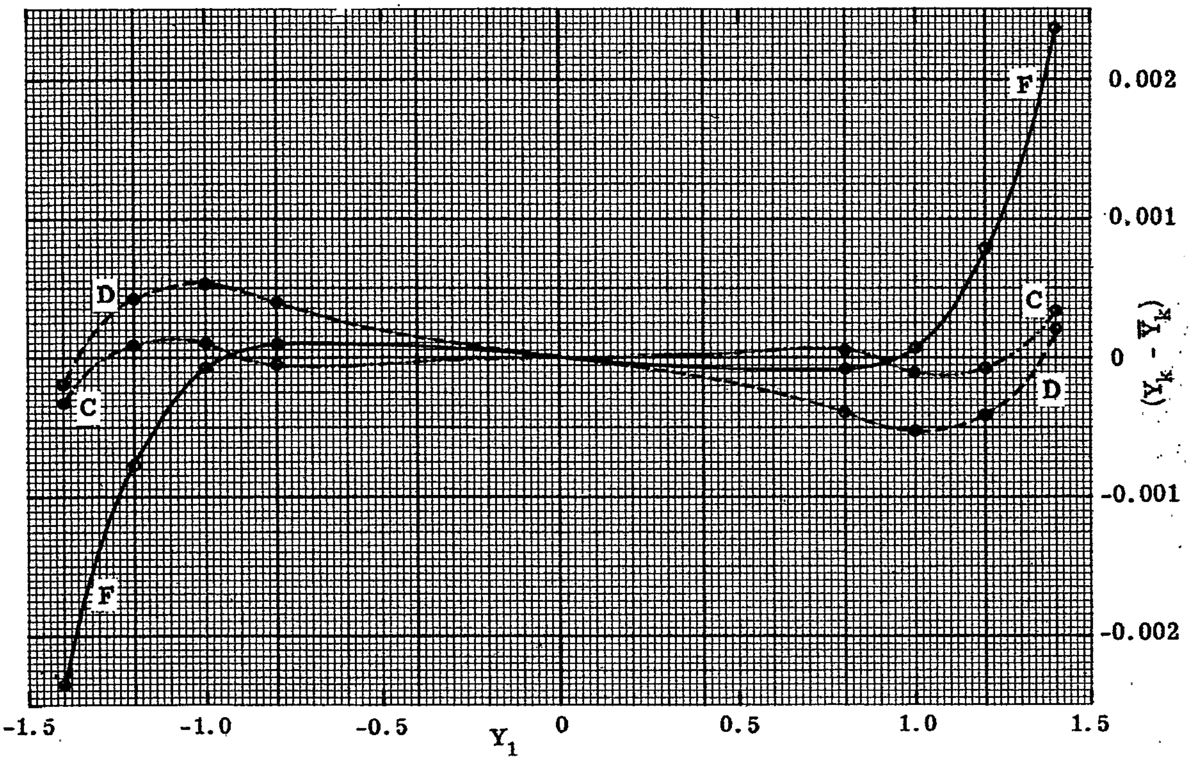


Figure 11.3 - Meridional ray plot at 0° for a doublet lens.

11.3.2.6 The third order calculations for this lens are included in Table 11.8. Inspection of these data gives a clue to why b_5 is positive. Surfaces 1, 2 and 4 have negative values of B. The only source of positive B is the 3rd surface. A surface always adds higher order aberration of the same sign as the third order. Since the 3rd surface introduces such a large positive third order contribution it over-balances the negative higher order contributions from the other surfaces. The result is a positive fifth order term. In designing the doublet it is considered advisable to adjust the third order aberration so that the higher order correction brings the curve back to $Y_k = 0$ for the rays at $Y_1 = Y_{1\,max}$. This leaves a residual aberration called zonal aberration. In figure 11.3 the zonal aberration for D light amounts to -0.000526.

c	0.16788	-0.24494	-0.24379	-0.07386	
t	0.50	0.01	0.30		
n	1.511	1.0	1.649		
y	1	0.97161	0.96954	0.95972	
u	-0.05677	-0.20739	-0.03274	-0.1	
B	-0.00106	-0.03273	0.03580	-0.00225	ΣB = -0.00023 $_3Y_k$ = -0.00115

Table 11.8 - Third order spherical contribution on each surface of the doublet shown in Table 11.6

11.3.3 Tolerance on zonal aberration.

11.3.3.1 The question of the tolerance on the zonal aberration cannot be explained fully here, but a few guide lines can be given.

(1) If the axial image must be corrected to be physically perfect it is necessary to reduce the zonal aberration to the following tolerance.

$$(Y_k)_{zone} = \frac{4.6\lambda}{L_{k-1}}$$

L_{k-1} = optical direction cosine of the emerging ray.

This tolerance assumes that the ray traced at a height of $Y_{1\,max.}$ is adjusted so that $Y_k = 0$. Tolerance is calculated as a positive number.

(2) If the objective is used in a telescope, one can compute the angular aberration presented to the eye, and set the tolerance by using the principle that the angular resolution of the eye is limited to one part in 3000. However, there is no need to attempt to reduce the zonal aberration below the value given above for the diffraction image case (1). In this region the size of the image is actually determined by the physical nature of the light and not by the geometrical aberrations.

11.3.3.2 The $Y_{1\,max.}$ for the lens shown in Table 11.6 is 1.4. The focal length is 10. Therefore, L_{k-1} is approximately -0.14 The zonal tolerance $(Y_k)_{zone}$ is then calculated as follows.

$$(Y_k)_{zone} = \frac{4.6 \times 0.5893 \text{ microns}}{0.14}$$

$$= 19.4 \text{ microns} = 0.0194 \text{ mm.}$$

The zonal aberration for the lens in Table 11.6 is -0.000526. This means that the lens could have a focal length 10 x 0.0194/0.000526 = 369mm and remain corrected within the tolerance. This assumes, of course, that the light is monochromatic. One can see that F and C light are not corrected as well as this. More will be said about this in a later section (Section 11.4) on secondary spectrum.

11.3.4 Methods for reducing the zonal aberration.

11.3.4.1 If the zonal aberration is too large in a lens it may be reduced by four methods. These methods will now be described for they illustrate a powerful technique of design. The methods are:

(1) Choosing the proper glasses.

(2) Using an air space.

(3) Introducing an aspheric surface.

(4) Adding an extra positive lens.

11.3.4.2 Tables 11.5, 11.6, and 11.7 illustrate the influence of glass choice.

11.3.4.3 If the air space is made larger the marginal rays have a chance to drop more before they strike the negative lens. The higher order negative aberration on the positive lens then causes the rays to actually strike the over-correcting surface at a lower aperture than predicted from first and third order theory. This cuts down on the higher order over-correcting tendency of this surface. Therefore as the space is increased the positive fifth order term is reduced. The third order value can then be made less negative, resulting in a reduced zone. Figures 11.4, 11.5, and 11.6 show some of the aberration curves for doublets where the air space has been adjusted to minimize the spherical aberration in D light. These lenses were also corrected so that the $Y_{1\,max.}$ was 1.4. The zonal aberration has been reduced to a remarkable degree. Table 11.9 contains the curvatures and thicknesses for many optimum solutions of this type. The last two columns are headed OSC', which stands for offense against the sine condition. This quantity, OSC', is proportional to coma, for a given image height, Y_k. These last two columns, therefore, are a measure of third order coma, and total coma for the marginal ray, respectively.

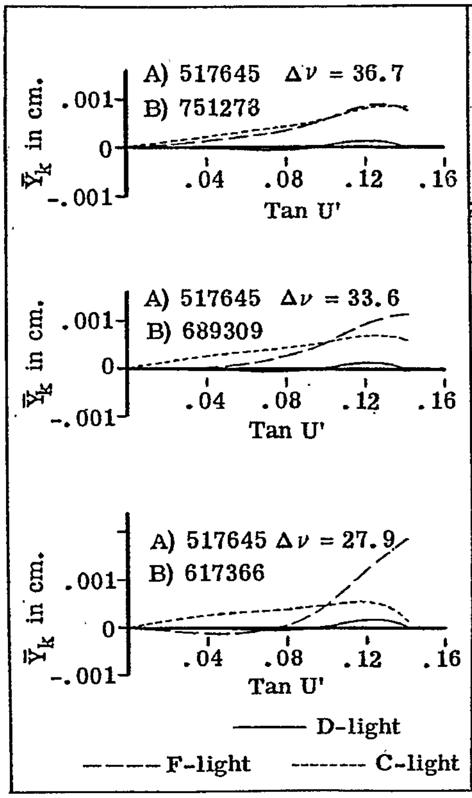


Figure 11.4- Effects of ν difference on spherochromatism (glass A used in positive lens, glass B used in negative lens).

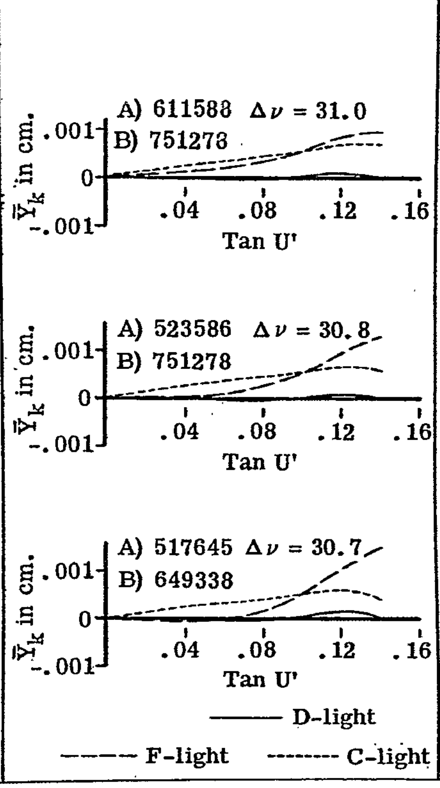


Figure 11.5- Effects of index of positive (A) lens on spherochromatism.

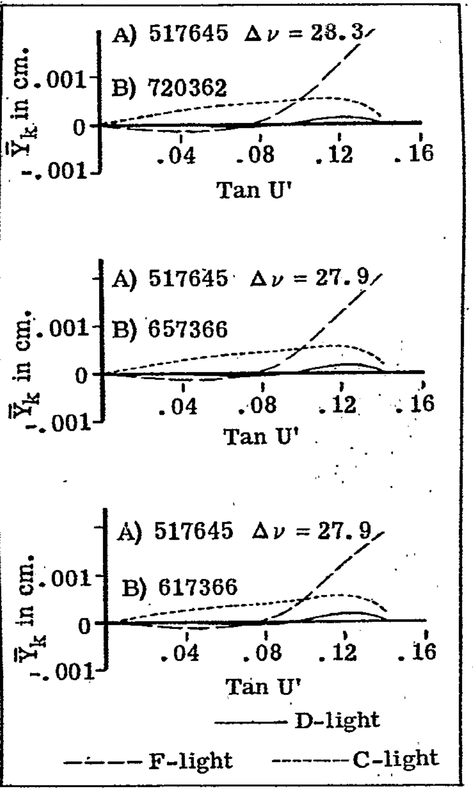


Figure 11.6- Effects of index of negative lens (B) on spherochromatism.

Glasses								OSC'	
+ lens	- lens	Δν	c_1	c_2	c_3	c_4	t_2	Third order	Marginal
517 645	751 278	36.7	+0.2582	−0.1450	−0.2222	−0.0489	+0.7364	−0.0030	−0.0017
517 645	720 293	35.2	+0.2379	−0.1630	−0.2159	−0.0470	+0.5353	−0.0026	−0.0014
517 645	689 309	33.6	+0.2205	−0.1825	−0.2173	−0.0468	+0.3596	−0.0022	−0.0012
517 645	649 338	30.7	+0.1955	−0.2215	−0.2348	−0.0516	+0.1412	−0.0015	−0.0009
517 645	617 366	27.9	+0.1780	−0.2638	−0.2663	−0.0575	+0.0400	−0.0007	−0.0006
517 645	657 366	27.9	+0.1634	−0.2716	−0.2671	−0.0783	+0.0022	+0.0003	−0.0000
517 645	720 362	28.3	+0.1526	−0.2733	−0.2632	−0.0986	−0.0261	+0.0010	+0.0004
523 586	751 278	30.8	+0.2091	−0.1801	−0.2067	−0.0593	+0.3141	−0.0016	−0.0009
523 586	720 293	29.3	+0.1955	−0.2016	−0.2161	−0.0612	+0.1826	−0.0013	−0.0008
523 586	689 309	27.7	+0.1829	−0.2260	−0.2317	−0.0641	+0.0883	−0.0008	−0.0006
523 586	649 338	24.8	+0.1611	−0.2776	−0.2730	−0.0751	−0.0014	+0.0004	+0.0000
523 586	617 366	22.0	+0.1373	−0.3417	−0.3334	−0.0926	−0.0221	+0.0028	+0.0015
523 586	657 366	22.0	+0.1149	−0.3787	−0.3653	−0.1282	−0.0311	+0.0048	+0.0026
523 586	720 362	22.4	+0.1009	−0.3636	−0.3478	−0.1523	−0.0444	+0.0064	+0.0039
611 588	751 278	31.0	+0.2149	−0.1381	−0.1858	−0.0111	+0.4904	−0.0020	−0.0011
611 588	720 293	29.5	+0.2052	−0.1551	−0.1913	−0.0086	+0.3540	−0.0019	−0.0011
611 588	689 309	27.9	+0.2006	−0.1731	−0.2027	−0.0030	+0.2606	−0.0020	−0.0011
611 588	649 338	25.0	+0.1971	−0.2091	−0.2308	+0.0079	+0.1481	−0.0025	−0.0015
611 588	617 366	22.2	+0.2161	−0.2426	−0.2662	+0.0389	+0.1128	−0.0050	−0.0032
611 588	657 366	22.2	+0.1713	−0.2598	−0.2652	−0.0142	+0.0347	−0.0013	−0.0009
611 588	720 362	22.6	+0.1439	−0.2676	−0.2623	−0.0538	−0.0128	+0.0008	+0.0003

Table 11.9- Final solution lens data resulting from least-squares correction program. For all lenses, $f' = 10.0$ cm, $t_1 = 0.5$ cm, $t_3 = 0.3$ cm.

11.3.4.4 Probably the simplest (from a theoretical point of view) method to reduce the zonal aberration is to introduce an aspheric surface on the last surface. Thus one can introduce high order terms of deformation and geometrically correct any amount of zonal aberration in the lens. The method used to compute the necessary coefficients is usually quite straight forward; briefly, it is as follows:

(1) Add a fourth order deformation term to reduce the third order aberration to zero. See Equation 8-(4a).

(2) Make an arbitrary guess at a sixth order coefficient (f). See Section 5.5.2.

(3) Trace through a ray at a finite aperture and determine how much of a deflection ΔY_k this aspheric term causes.

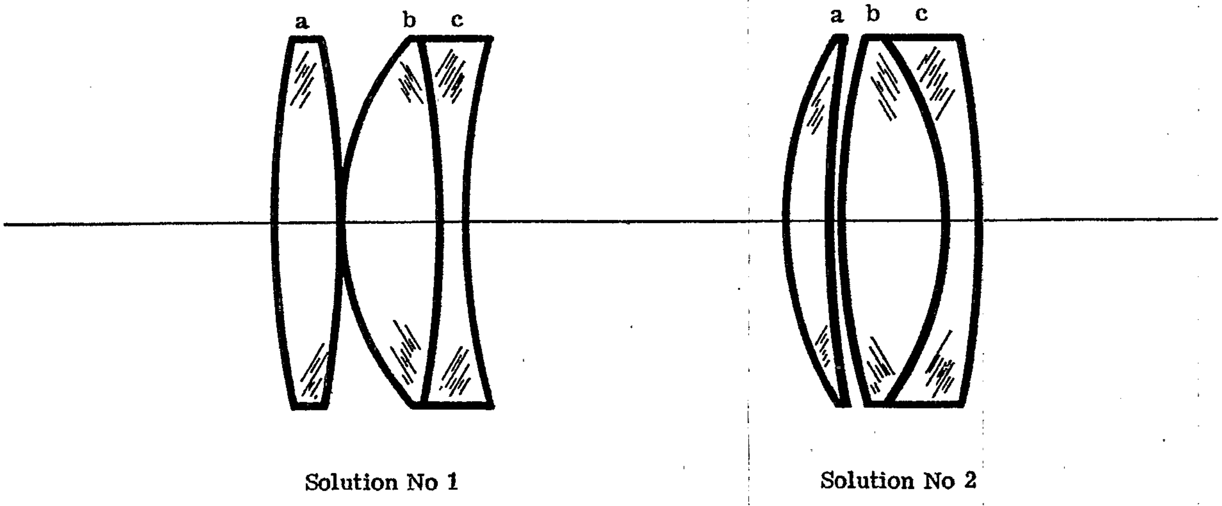
(4) It may then be assumed that this deflection

$$\Delta Y_k = \lambda \left[6fS^5 + 8gS^7 + 10hS^9 + \ldots\ldots \right] .$$

(5) It is then possible to write a set of equations to bring as many rays to the axis as there are aspheric constants. It is possible to write as many equations as rays traced through the system, but if there are more equations than terms in the aspheric, one has to resort to a method of least squares rather than expect an exact solution.

(6) Since the equation in step 4 is not exact, it may be necessary to repeat the process a few times.

This method is usually satisfactory, but if either the aperture of the lens or the zonal aberration is large, it may not be possible to fit a power series deformation which will reduce the aberration for all rays. This is because not enough terms are used in the expansion. In practice if one cannot fit a curve with a 10th degree polynomial then it helps very little to add a few more terms in the series; it takes a large number of terms to reduce the aberration for many rays. Sometimes it becomes necessary to abandon the use of the polynomial expression. The aspheric must then be expressed as a series of Y and Z coordinates. This is computed by actually calculating the optical path along the ray, and adding glass thickness to produce a spherical wavefront. This procedure is almost never necessary, but if it is, then one seriously questions whether it would be possible to make an aspheric of this type.



Solution No 1 Solution No 2

Figure 11.7 - Two types of triplet solution.

11.3.4.5 The zonal spherical aberration can be reduced by splitting off some of the power in the positive lens in the doublet. This would make a triplet similar to the types illustrated in Figure 11.7. By so doing, two extra degrees of freedom are created, namely the power ϕ of the extra positive lens, and its first curvature, c_1 . One degree of freedom will be eliminated by cementing the last two elements. The problem then becomes, what is ϕ_a ? In other words, how should ϕ_a and ϕ_b be distributed? The manner one uses to find an answer to this problem is typical of one of the techniques used by a lens designer. The reasoning goes as follows:

(1) If the lens is assumed to be thin, the power $\phi_a + \phi_b$ will equal the same power as for the positive lens in a doublet. Therefore one may start by using the powers for the lenses as given in Table 11.3.

(2) By cementing the b and c lenses, there are only two degrees of freedom left. The first curvatures, c_1 and c_3 , are the variables. If ϕ_a is decided upon, then the thin lens formulae, described in Section 8.9, enable one to compute the coefficients of the equations

$$B_a = \alpha_{1a} + \alpha_{2a} c_1 + \alpha_{3a} c_1^2 ,$$

$$B_{b+c} = \alpha_{1(b+c)} + \alpha_{2(b+c)} c_3 + \alpha_{3(b+c)} c_3^2 ,$$

$$F_a = \beta_{1a} + \beta_{2a} c_1 ,$$

$$F_b = \beta_{1b} + \beta_{2b} c_3 .$$

In exactly the same way as for the doublet, two types of solutions may be found. They are illustrated in Figure 11.7.

(3) Next plot B_c for these two solutions on a plot as shown in Figure 11.8. B_c is the total spherical aberration due to the negative lens. By finding the two solutions for several values of ϕ_a it is possible to plot the curves shown in Figure 11.8. These curves show that if $\phi_a = 0$ we then have a simple doublet and the two solutions require different amounts of positive spherical aberration. For the doublet, of course, the (b) and (c) lenses must be considered to be separated. As more and more power is put into lens (a), less and less positive B is required of lens (c). Now we know that the higher order aberrations will be minimized when the lens is corrected with B_c having a minimum positive value because the higher order spherical aberration has the same sign as the third order. From this reasoning one would predict that the type 1 solution with a value of $\phi_a = 0.066$ would provide an optimum solution. Solution 1, shown in Figure 11.7, is a lens of this type. The type 2 lens was also chosen with a value of $\phi_a = 0.066$.* Table 11.10 shows the results of ray tracing these solutions after adjustments were made to the residual third order aberration so that the marginal ray comes to focus at $Y_k = 0$. Table 11.11 contains the data for these two solutions.

(4) It is interesting to see that the type 1 lens is remarkably well corrected. The zonal aberration is 10 times less than the type 2. The type 2 lens may be thought of as a derivative of a separated doublet of the left hand branch. With the choice of glass used, the doublet would be an air spaced lens. By cementing it, it would be under-corrected for spherical aberration. Now by splitting off a small part of the positive lens and by bending slightly the lens can be re-corrected for spherical aberration, thus leading to the lens type 2. The type 1 lens is actually a derivative from the right hand branch of the doublet. Note that the better solution comes from the poorer doublet type. This is mentioned because, in designing this type lens, if one started by trying to modify a left hand doublet lens he might easily converge on a type 2 solution and find no advantage in using the split positive lens. Notice that the zonal aberration in the type 2 lens is

---

* It is true that the type 2 solution would probably be better at $\phi_a = 0.082$ or at $\phi_a = 0.138$, but the value of $\phi_a = 0.066$ was selected to illustrate that for a given value of ϕ_a there are two solutions quite close together.

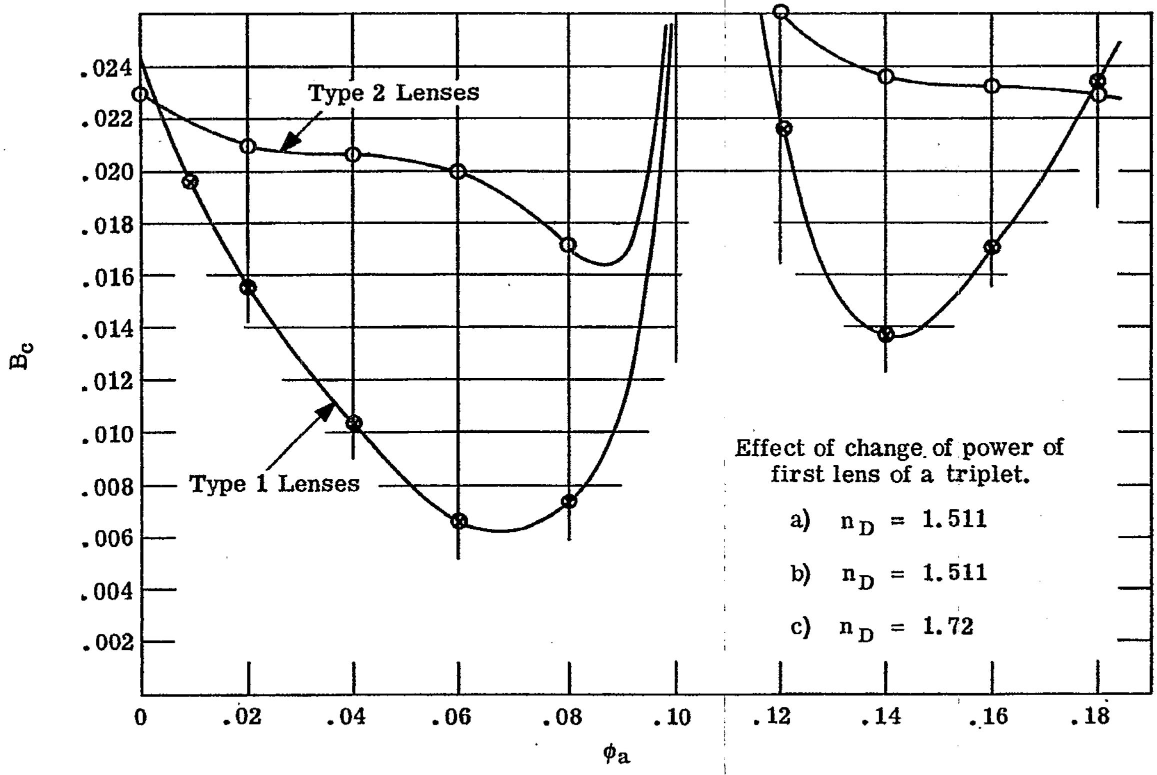


Figure 11.8 - A plot of B_c the total spherical aberration of the negative lens as a function of ϕ_a .

Y_1	TYPE 1 Y_k	TYPE 2 Y_k
1.4	0.000210	-0.000101
1.2	0.000070	-0.000706
1.0	0.000010	-0.000696
0.8	-0.000007	-0.0004681

Table 11.10 - Ray trace data in D light showing a comparison between type 1 and type 2 solutions in triplet telescope objectives

Type 1				Type 2	
c	t	n_D	ν	c	t
0.07121				0.1641	
	0	1.511	63.5		0
-0.05525				0.0349	
	0				0
0.20132				0.0472	
	0	1.511	63.5		0
-0.03572				-0.1871	
	0	1.72	29.3		0
0.08337				-0.0680	
		f' =	10		

Table 11.11 - Lens data on type 1 and type 2 lenses

almost identical with the doublets shown in Table 11.6. If we had happened to choose a glass combination which would have resulted in a cemented doublet, then a type 2 solution would offer no advantage. One would have to go to type 1. The curves shown in Figure 11.8 change as the glass is varied. In Figure 11.9 the type 1 branch is shown for another pair of glasses. One can see that it is quite different, for most of the positive power should be placed in the (a) lens.

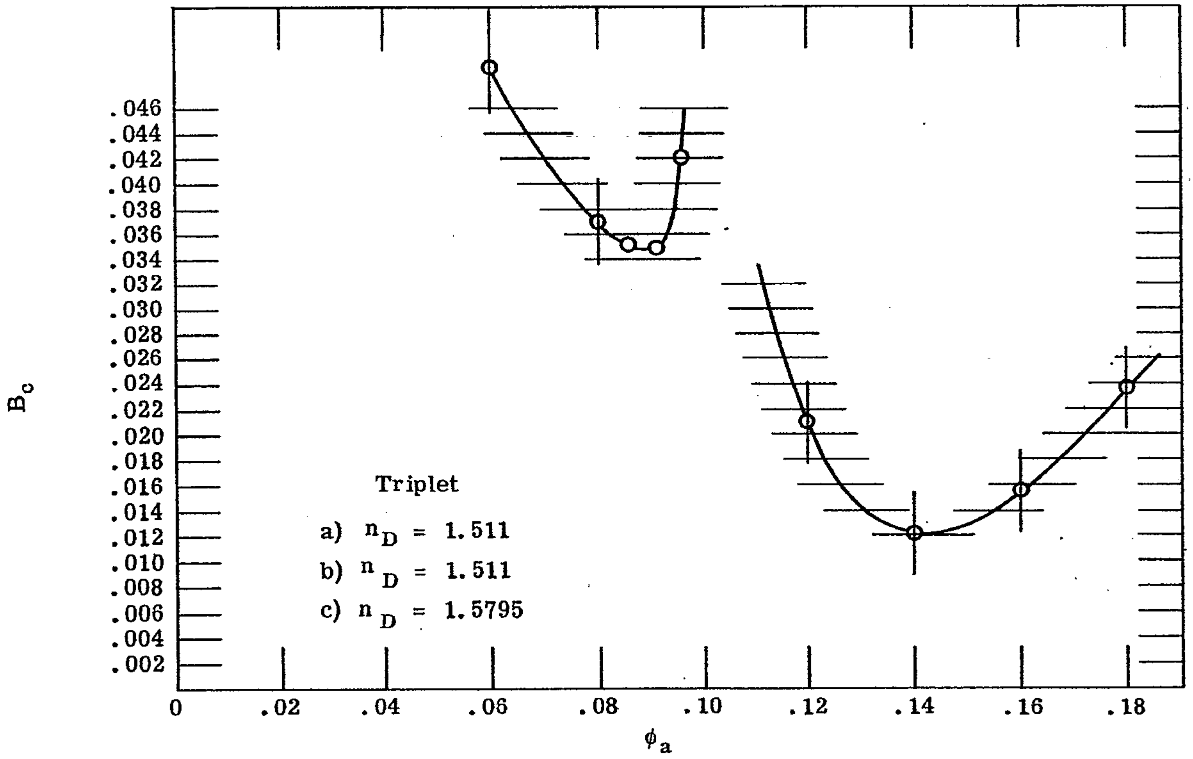


Figure 11.9 - A plot of B_c as a function of ϕ_a for a triplet.

11.3.5 Discussion of zonal correction methods.

11.3.5.1 This study illustrates that there are many possible solutions, and the designer must investigate them all in order to be sure he has exhausted the possibilities. This study also illustrates that there are regions where no solutions exist even though there appear to be a sufficient number of degrees of freedom.

11.3.5.2 If possible, a doublet design should be used. However if the zonal spherical aberration exceeds the tolerance, there are the four methods for reducing the zonal aberration described in the above section. Of these methods the fourth method of adding the extra element is strongly recommended, for this method does not involve the balance of large aberrations. The other three methods depend upon the balancing of large aberrations. The extra element method provides a solution with no large surface contributions. One will find in practice that it also will provide a lens much easier to manufacture, for it will be much less sensitive to decentering or spacing errors.

11.3.6 Coma correction. So far practically nothing has been said about coma correction. All the doublet solutions were corrected to have the third order coma exactly zero. This provides optimum correction for most of the designs. However, if the zonal spherical aberration is corrected by a large air space, one will encounter high order coma aberration, and it may be necessary to introduce residual coma in the third order. This however is very unsatisfactory, so everything possible should be done to find a solution with minimized high order coma. For this reason also, the split doublet lens offers a far better way to reduce the zonal aberrations than does the unsplit doublet, as the former has excellent coma correction.

11.4 SECONDARY SPECTRUM OF TELESCOPE OBJECTIVES

11.4.1 The difference in focus for F, C, and D light.

11.4.1.1 In the preceding paragraphs, it was shown (11.3.4.3) how the zonal spherical aberration could be reduced by a large factor. However, the aberration curves of Figures 11.4 through 11.6 clearly show that since the lens must be designed to image F and C light, as well as D light, the high degree of zonal correction in D light is of small practical significance. The F and C light focus does not coincide with the D focus. This defect in focus for F, C, and D light is a paraxial ray defect and was briefly discussed in Section 6.10.8, where the transverse aberration, $TAch_{F-D}$, was defined as the secondary spectrum.

11.4.1.2 It was also stated (6.10.8.4) that the three principal methods for reduction of this aberration are:

(1) Use special materials with equal partial dispersions.

(2) Use more than two types of glass.

(3) Use proper combination of lenses.

Paragraphs 11.4.2 and 11.4.3 will describe methods (1) and (2) respectively.

11.4.2 Reduction of secondary spectrum in a doublet. ($\tilde{P}_a - \tilde{P}_b = 0$ Method).

11.4.2.1 Equation 6-(49) tells us that when the partial dispersion ratios of both lenses of a doublet are equal for F and D light, then $\tilde{P}_a - \tilde{P}_b = 0$ and the F, C, and D light will unite in a common focus. Now, depending on the shapes of the dispersion curves of the glasses used in the doublet, there is still the question of where other wavelengths will focus.

11.4.2.2 Equation 6-(49) can also be used to calculate the $TAch_{\lambda-D}$ for any other wavelength. If the partial dispersion ratios for other wavelengths are not equal, i.e., $\tilde{P}_a - \tilde{P}_b \neq 0$, then there is still residual chromatic aberration. In choosing glass types, then, a designer must consider the following compromises:

(1) Should he settle for a small $(\nu_a - \nu_b)$ by setting $(\tilde{P}_a - \tilde{P}_b)$ exactly equal to zero, or should a larger $(\nu_a - \nu_b)$ be chosen and some secondary color be allowed? The decision, of course, will depend on the focal length and the numerical aperture required of the objective.

(2) Does the need for correction for a large range of wavelengths require that $(\tilde{P}_a - \tilde{P}_b)$ be set at a value other than zero?

11.4.2.3 It is clear that these considerations, combined with the task of correcting the spherical aberration, and the variation of spherical aberration with wavelength, pose a formidable array of problems.

11.4.2.4 The designer's difficulties are further increased by the need for extremely accurate measurements of the index of refraction. One can, by differentiating Equation 6-(49), determine that the following relation holds for an achromatic doublet.

$$dn_\lambda = \frac{d\left(TAch_{\lambda-D}\right)}{TAch_{\lambda-D}} \left(\frac{n_D - 1}{2200}\right) \qquad (13)$$

Therefore, if it is desired that the secondary spectrum of a doublet be held to 1/10 of its normal value, then it should be sufficient to know the index of refraction at each wavelength with an accuracy of 2 in the fifth decimal place. With an index error of this magnitude in each of the wavelengths used to calculate $\tilde{P}$, and ν, it is possible for the errors to combine so as to cause a doubling of the total error. Thus, it is necessary that the index be accurate to at least half of this value, or 1.0 in the fifth decimal place.

11.4.2.5 It is not only difficult to make measurements of the index of refraction with this accuracy; it is even more difficult to manufacture glass to specification with this degree of precision. The reputable optical glass manufacturers claim the required accuracy of their measurements but do not claim to be able to furnish samples to catalog values with this exactness. If, in the manufacture of precision lenses, it is necessary to have glass whose index of refraction is accurate to this degree, then it is necessary to have measurements of index made on a sample of the actual glass to be used in the lens.

11.4.3 Correction of secondary spectrum in a triplet lens. (Multiple glass-type method).

11.4.3.1 By using three glass types in the telescope objective it is possible in principle to bring at least three wavelengths to a common focus. If it is assumed the lenses are all thin and closely spaced then it is possible to write the following equations.

$$\phi_a + \phi_b + \phi_c = \phi \qquad \text{(Focal length)} \qquad (14)$$

$$\frac{\phi_a}{\nu_a} + \frac{\phi_b}{\nu_b} + \frac{\phi_c}{\nu_c} = 0 \qquad \text{(F and C light brought to same axial focus)} \qquad (15)$$

$$\frac{\phi_a P^*_a}{\nu_a} + \frac{\phi_b P^*_b}{\nu_b} + \frac{\phi_c P^*_c}{\nu_c} = 0 \qquad \text{(D light brought to the F-C axial focus)} \qquad (16)$$

11.4.3.2 By defining $\tilde{P}^* = \dfrac{n_F - n_D}{n_F - n_C}$, the third equation must be fulfilled if D light is to be focused at the F and C focus. The above equations may be solved to give the following values of ϕ_a , ϕ_b , and ϕ_c .

$$\phi_a = \phi \frac{\nu_a \left[\tilde{P}^*_c - \tilde{P}^*_b \right]}{\Delta} , \qquad (17)$$

$$\phi_b = \phi \frac{\nu_b \left[\tilde{P}^*_a - \tilde{P}^*_c \right]}{\Delta} , \qquad (18)$$

$$\phi_c = \phi \frac{\nu_c \left[\tilde{P}^*_b - \tilde{P}^*_a \right]}{\Delta} , \qquad (19)$$

where

$$\Delta = \begin{vmatrix} \tilde{P}^*_a & \nu_a & 1 \\ \tilde{P}^*_b & \nu_b & 1 \\ \tilde{P}^*_c & \nu_c & 1 \end{vmatrix} . \qquad (20)$$

One can recognize that Δ is a determinant, the value of which is the area of a triangle connecting points plotted with $\tilde{P}^*$ as the ordinate and ν as the abscissa.

11.4.3.3 Figure 11.10 is such a plot for several glasses. From the above equations one can see that points for three glasses must be found so as to form a triangle of finite area on the $\tilde{P}^*$ versus ν plot. It is important to pick glasses that will have the smallest possible values of ϕ_a , ϕ_b , and ϕ_c . If three glasses are picked, as shown in Figure 11.10 marked as a, b, c, then the (b) lens becomes negative, for $(\tilde{P}^*_a - \tilde{P}^*_c)$ is negative and Δ is positive. In order to minimize ϕ_b , the ratio of $(\tilde{P}^*_a - \tilde{P}^*_c)/\Delta$ must be made a minimum. If one draws a line from a to b it is clear that any glasses located on this line will have the same ratio of $(\tilde{P}^*_a - \tilde{P}^*_c)/\Delta$. If $\tilde{P}^*_a - \tilde{P}^*_c$ is made smaller, the area Δ of the triangle is made smaller by the same ratio. This can be seen to be true by remembering that ac is the base of the triangle and a perpendicular from b to this base line is the altitude of the triangle. Therefore,

$$\frac{(\tilde{P}^*_a - \tilde{P}^*_c)}{\cos\theta} h = \Delta ,$$

and

$$\left[\frac{\tilde{P}^*_a - \tilde{P}^*_c}{\Delta} \right] = \frac{\cos\theta}{h} .$$

$\cos\theta$ is the angle between the line connecting a and c and the vertical axis. As long as $\cos\theta$ and h remain constant then $(\tilde{P}^*_a - \tilde{P}^*_c)/\Delta$ is constant. The procedure to pick glasses, then, is to try to find a triangle with as large an h as possible. It is also logical to suggest that the two positive lenses [the (a) and the (c) lenses] should have approximately the same power.

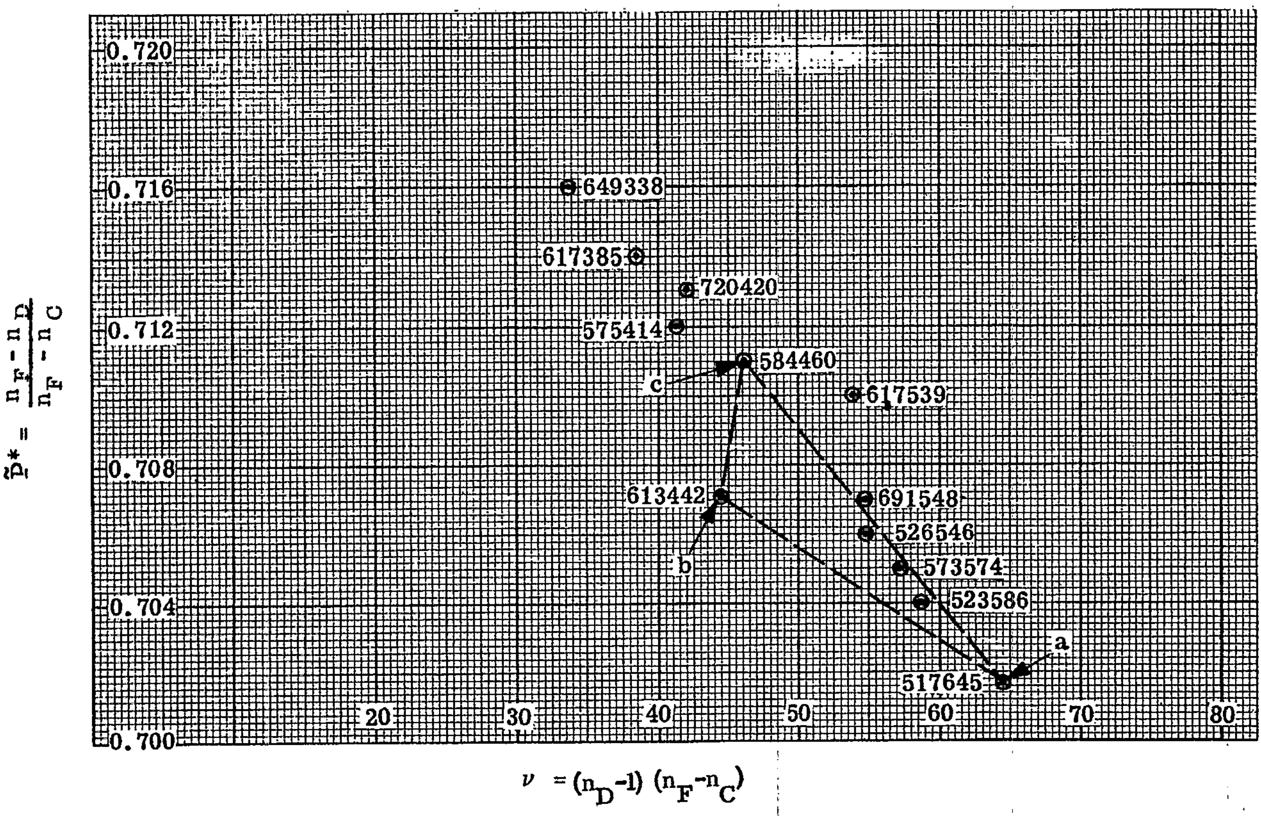


Figure 11.10- A plot of $\tilde P^*$ vs ν for different glasses.

11.4.3.4 Dividing equation (17) and (19) provides the ratio ϕ_a / ϕ_c .

$$\frac{\phi_a}{\phi_c} = \frac{\nu_a}{\nu_c} \left(\frac{\tilde P^*_c - \tilde P^*_b}{\tilde P^*_b - \tilde P^*_a} \right)$$

Since ν_a / ν_c is greater than 1, it then follows that $(\tilde P^*_c - \tilde P^*_a)$ should be less than $(\tilde P^*_b - \tilde P^*_a)$. A glass combination selected with these considerations is shown in Figure 11.10. There are, however, other factors one must consider in selecting the glasses for the reduction of secondary color, namely, tertiary color as described below.

11.4.3.5* By satisfying the conditions in Equations (14), (15) and (16), the three wavelengths F, D, and C focus at a common axial point. One may now calculate the residual transverse aberration for any other wavelength λ , from the equation

$$\left(\frac{\phi \tilde P_{\lambda - D}}{\nu} \right)_a + \left(\frac{\phi \tilde P_{\lambda - D}}{\nu} \right)_b + \left(\frac{\phi \tilde P_{\lambda - D}}{\nu} \right)_c = \mathrm{TAch}_{F-\lambda} \left(\frac{n_{k-1}\ u_{k-1}}{y^2} \right) \qquad (21)$$

$\tilde P_{\lambda - D}$ will be hereafter referred to as $\tilde P^{**}$. By inserting the expressions for (ϕ / ν) given in Equations (17), (18), (19), and (20), Equation (21) becomes

$$\frac{\tilde P^{**}_a\ (\tilde P^*_c - \tilde P^*_b) + \tilde P^{**}_b\ (\tilde P^*_a - \tilde P^*_c) + \tilde P^{**}_c\ (\tilde P^*_b - \tilde P^*_a)}{\Delta} = \mathrm{TAch}_{F-\lambda} \frac{n_{k-1}\ u_{k-1}}{y^2\ \phi} . \qquad (22)$$

---

* The notation P* and P** and the ideas suggested in this section have been described by Herzberger, Optica Acta, 6, 197 (1959).

The left hand side of the equations is equal to

$$\frac{\begin{vmatrix} \tilde{P}^*_a & \tilde{P}^{**}_a & 1 \\ \tilde{P}^*_b & \tilde{P}^{**}_b & 1 \\ \tilde{P}^*_c & \tilde{P}^{**}_c & 1 \end{vmatrix}}{\Delta}$$

The value of the determinant in the numerator is again the area of a triangle in a coordinate system with P^* plotted as abscissa and P^{**} plotted as ordinate. This tells us then, that if we wish to have small residual aberration for other wavelengths it is necessary to pick three glasses that lie on a straight line when plotted on the P^* versus P^{**} diagram. Plots of this type are shown in Figure 11.11. There are three sets of wavelength data plotted on this graph. The values of P^{**} are $P_{A'-D}$, P_{e-D}, and P_{g-D}. The glasses used in a sample calculation are shown connected by dotted lines. These plots show that A', g and e light will have residual aberration because the three glasses do not lie on a straight line. As the data is plotted the triangles show that A' will have a positive TAch. The e light will be slightly negative, and g will be slightly positive. An actual curve for these three glasses is shown in Figure 11.12. It is plotted on the same coordinates as the data in Figure 6.20. The corresponding curve for a doublet is shown in the same figure. The powers of the lenses in the triplet are shown compared with a doublet in Table 11.12. The strong curves in the triplet indicate clearly the reason why lenses corrected for secondary color must have small relative apertures.

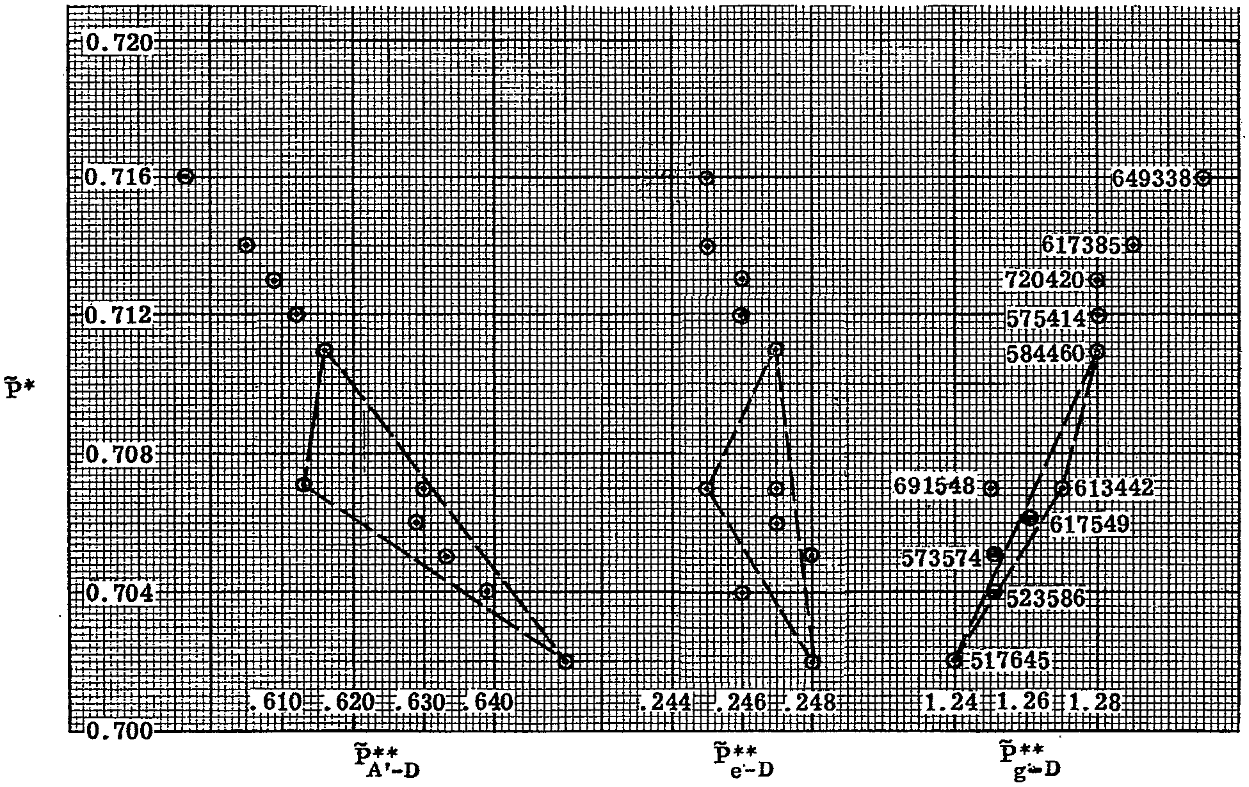


Figure 11.11 - A plot showing tertiary spectrum.

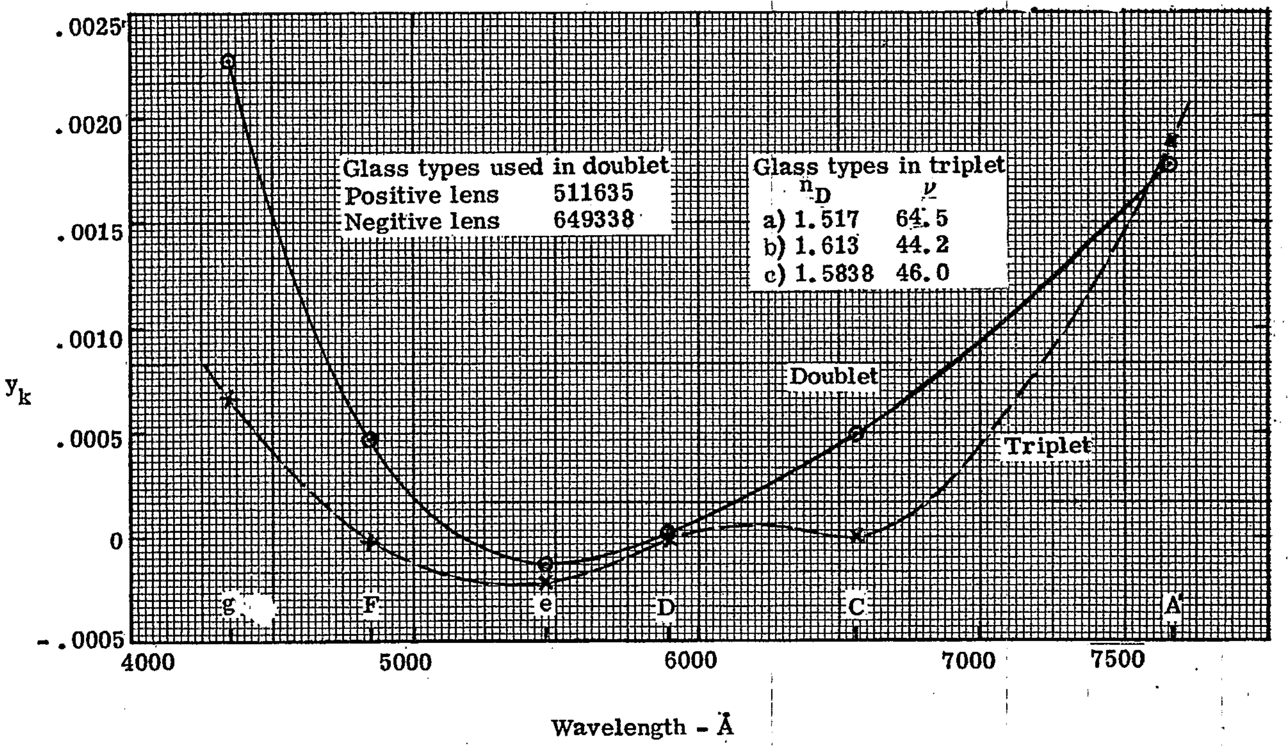


Figure 11.12 - Plot of y_k vs λ for a triplet corrected for secondary color.

Doublet	Triplet
ϕ_a = 0.21380	ϕ_a = 0.2812
ϕ_b = -0.1138	ϕ_b = -0.4747
	ϕ_c = 0.2935
ϕ = 0.1	ϕ = 0.1

Table 11.12 - Comparison between powers in a triplet corrected for secondary color and an ordinary doublet.

11.4.4 Additional readings on secondary color. For further reading on secondary color, refer to the following articles:

(a) Three color achromats, R. E. Stephens, J. Opt. Soc. Am. 49, 398 (1959)

(b) Four-color achromats and superchromats, R. E. Stephens, J. Opt. Soc. Am. 50, 1016 (1960)

11.4.5 Sample design of a triplet corrected for secondary color.

11.4.5.1 A sample lens has been fully corrected for secondary color and ray traced. The glasses used in the design were all Schott glasses. The thin lens solution was given by R. E. Stephens in the second of the above papers. The thin lens glasses and powers were given as follows:

Glass	Power
F-1	0.338
KzFS-4	-0.721
PKS-1	0.483

These powers add up to 0.1. The focal length is therefore 10. The lens was corrected for an f-number of 12. The final lens specifications are shown in Table 11.13.

LENS SPECIFICATIONS

c	t	Glass
0.1989	0.1683	F-1
-0.0791	0.0175	Air
-0.1502	0.1400	KzFS-4
0.6180	0.0100	Air
0.6387	0.2964	PKS-1
-0.1272	9.2728	Air

RAY TRACE DATA

Y_1	$(Y_k)_D$	$(Y_k)_F$	$(Y_k)_C$
0.25	0.00000369	0.000092	0.000023
0.375	-0.000155	0.000074	-0.000164
0.500	-0.001056	-0.000574	-0.001137

Table 11.13 - Three lens system corrected for secondary color.

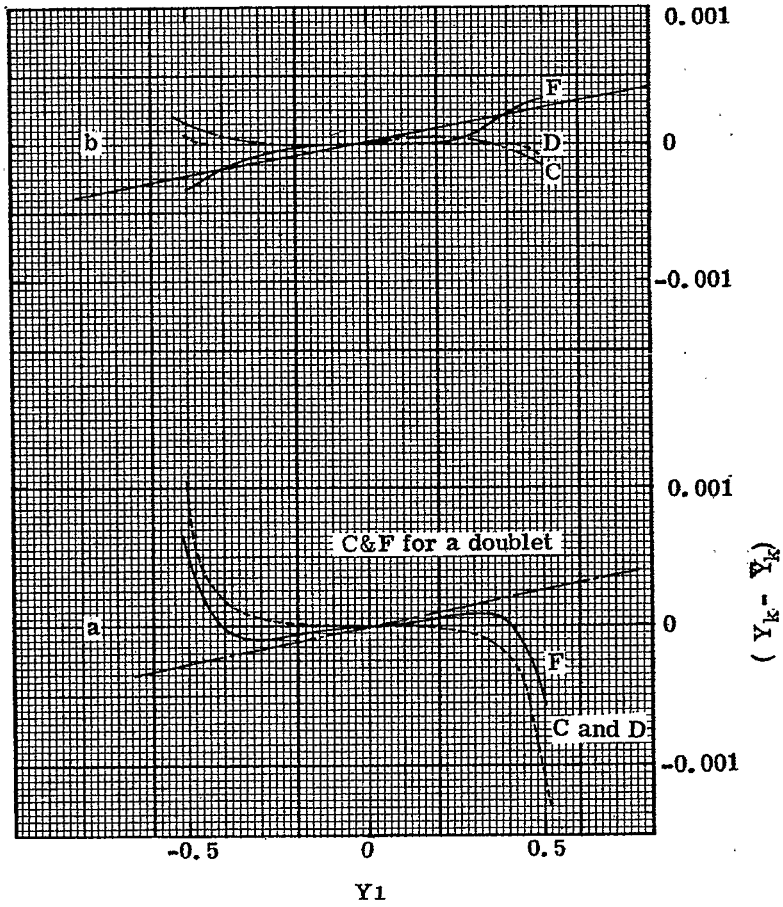


Figure 11.13 - Meridional ray plot at 0°, for triplet.

The meridional ray plot for this lens is shown in Figure 11.13a. It is plotted on the same scale as the doublet shown in Figure 11.3. These data show that the triplet corrected for secondary color has to be made to a much larger ƒ-number than the doublet. One can see that the curves are stronger than the doublet, and the higher order spherical aberration is large. The straight dashed line in Figure 11.13a shows the best possible aberration curves for F and C light in a simple doublet. The curves for the triplet out to a value of $Y_1 = 0.4$ show that some advantage is gained by using the triplet. Beyond $Y_1 = 0.4$ there is no gain at all, for the high order spherical aberration is so heavily under-corrected. If the lens had been stopped down to a value of $Y_1 = 0.3$ it could have been corrected with a smaller residual aberration. The lens would then be corrected with approximately one third the aberration in a simple doublet. This would give an ƒ-number of 16.7.

11.4.5.2 Other solutions. The triplet described in Table 11.13 is not a completely optimized solution. The higher order aberrations might have been further reduced by adjusting the air spaces. One must also consider other orientations of the lenses. To illustrate this effect a solution was corrected with the PKS-1 as the first element, and F-1 as the final element. This solution is shown in Table 11.14. This shows some improvement over the first solution shown. This lens could probably be used at ƒ/11. This lens is better corrected than the one with F-1 in front. It is not, however, certain that this is a characteristic of the lens. The second system was designed much more carefully than the first one. Many, many solutions were found for the second design. The solutions were found automatically for varying amounts of primary and secondary color, and finally, the zonal spherical aberration was reduced by adjusting the air space between the second and third lens. The second solution is, we believe, nearly as good a solution as it is possible to design with this combination of glass; but we are not sure, for there are many things that should be studied. For example, the axial color should probably be made slightly more negative. This would lower the F light curve slightly and raise the C light curve. The two curves would therefore cross further out in the aperture.

LENS SPECIFICATIONS

c	t	Glass
0.5334	0.3180	PKS-1
-0.3322	0.0100	Air
-0.2792	0.1400	KzFS-4
0.7079	0.1000	Air
0.5469	0.1969	F-1
-0.1723	9.018	Air

RAY TRACE DATA

Y_1	$(Y_k)_D$	$(Y_k)_F$	$(Y_k)_C$
0.25	0.0000021	0.0000127	0.0000517
0.375	0.0000028	0.0001495	0.0000265
0.500	-0.000085	0.0003700	-0.0001531

Table 11.14- Three lens system corrected for secondary color.

The meridional ray plot is shown in Figure 11.13b.

11.4.5.3 The amount of computing that went into the above designs is beyond the comprehension of anyone not familiar with the problem. We found 16 automatic third order solutions. Usually it took a minimum of four iterations. For each solution a fifth order and 9 rays were traced. Only one out of five possible thicknesses was used as a variable. Before one could say he really had an optimum solution it would be necessary to check the effectiveness of varying the thickness, trying the negative lens out front and use other glasses. Thanks to the modern computer it is beginning to become practical to do this at a reasonable cost. When we realize how limited our present approaches are to the problem, we can look forward to promising solutions in the future.

11.4.6 Evaluation of lens from optical path. The gain in image quality is however, misleading. This is an example of why one must always remember to consider the physical optics of the problem. Figure 11.13 shows that if the perfect doublet is stopped down to $f/16.7$ the C and F light would have a transverse aberration of 0.00015, with respect to the D light focus. If we assume the focal length is 10 cm then this corresponds to a transverse aberration of 0.00015 cm. There is a relation between the transverse aberration and optical path difference for shift of focus. The equation is

$$\mathrm{OPD} = Y_k \cdot \frac{Y_1}{2 f'}$$

where Y_k is the transverse aberration for the ray entering the lens at Y_1. Inserting the above aberration into this equation shows that the OPD in the simple doublet is 0.038 wavelengths. Now if the Rayleigh tolerance of $\lambda/4$ is assumed, this means an $f/16.7$ doublet with a focal length of 10 cm has so little secondary spectrum it will never be noticed. Its focal length could be scaled up 0.25/0.038 times the 10 cm focal length. This amounts to a focal length of 65.8 cm. Therefore, there is no point whatsoever in designing a lens to reduce the secondary spectrum, when the focal length is less than 65.8 cm, if it has to be stopped down to $f/16.7$. The above triplet therefore would not show any advantage until scaled to focal lengths longer than 65.8 cm. To realize a two to one gain over a doublet, it will have to be scaled to 131.6 cm focal length. At $f/16.7$ this would be a lens 7.9 cm in diameter. This is getting to be a fairly expensive size lens in which to use the unusual glass KzFS-4.

11.5 SUMMARY

One can now see the relation between a doublet and a triplet corrected for secondary color. One can design a doublet with two types of glasses and split the positive lens into two and make the system a triplet. Then it is necessary to find two glasses with the same values of P^* and P^{**}. Since there are only very few glasses removed from the $\tilde{P}$ versus ν line this means that relatively few glasses are available for the positive lens. By using a third glass type it is possible to use a much larger selection of glasses. In order to find an optimum solution it is necessary to study many combinations of glass. One must completely correct the lens and ray trace the solution before deciding what glass choices are most suitable. There will be variation of spherical aberration with wavelength. This aberration may become so large that all the advantage of using the special glasses to correct the first order effects may be completely lost.

12 LENS RELAY SYSTEMS

12.1 INTRODUCTION

12.1.1 Lens relay systems of the type mentioned in Section 7 (Paragraph 7.5 and Figures 7.5 and 7.6) are described in more detail here.

12.1.2 Relay systems are used for two purposes: to provide the proper orientation of the image, and to transfer the light from one region to another. Sometimes the distance between the object and the final image may be large and, in addition, the diameter of the lenses must not become excessive. These conditions may require a series of relay lenses resulting in a system called a periscope.

12.2 THE BASIC LENS PROBLEM OF A RELAY SYSTEM

12.2.1 Suppose a relay system is needed to transfer light from an object plane to an image plane. The two planes are separated by the distance D, Figure 12.1. The magnification should be -1.

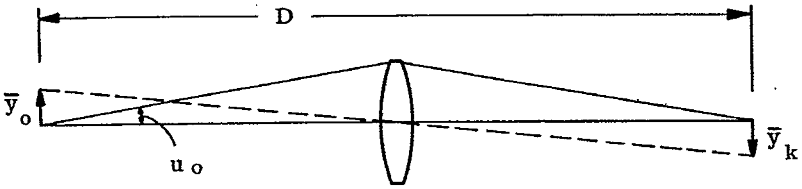


Figure 12.1 - A single-relay-system.

12.2.2 The diameter of the objective will be determined by the angle u_o. The type of lens to use for the objective depends upon the image quality required. For the moment, however, assume that the relay lens will consist of two telescope objectives of the type described in Section 11. The objectives could be placed so that the light would be parallel between them. One could start with any of the lenses shown in Tables 11.3, 11.5, 11.6 or 11.7.

12.3 A VISUAL SYSTEM, NUMERICAL EXAMPLE

12.3.1 To evaluate the visual performance with telescope objectives, first suppose that the system is a visual system, with a 10X eyepiece (see Section 14) used to view the image.

12.3.2 In Section 11.1.3 the effect of the Petzval curvature was described. Equation 8-(28) gives the value of P for a thin lens as $-\phi/n$. The value of P for the relay lens shown in Figure 12.1 would then be $-\frac{4}{Dn_e}$, where n_e is the effective index of the doublets used for the relay lens. The 10X eyepiece could be any of those described in Section 14. For this example, the very common Erfle eyepiece shown in Figure 14.19 will be used. According to Table 14.7, the lens will have a value of $P = -0.2125$ in reciprocal cm.

12.3.3 From the data shown in Section 11 on doublets, it is evident that the doublets could be used at f/3.5. If they could be 10 cm in diameter, each doublet could have a focal length of 35 cm, and the distance D would be approximately 70 cm. The value of P for the relay lens would then be -0.038.

12.3.4 This is only 18% of the Petzval contribution introduced by the eyepiece. It is, therefore, probably negligible as long as the field of view is maintained within the field of the eyepiece. The data in Figure 14.21 show that this eyepiece can cover about 28° half field with a negative distortion of approximately 8%. The maximum image height is therefore 1.32 cm. This means that the maximum object height will be 1.32 cm.

12.3.5 The use of other types of lenses for relay objectives is considered here. If two triplet objectives of the type designed in Section 10 were used, the value of n_e in the equation $P = -\phi/n_e$ could be raised to about 3.0 or 4.0. With a value of $n_e = 4.0$, the P value of the relay lens would be -0.014. This would be completely negligible compared to the value introduced by the eyepiece. On the other hand, the triplet would not be as well corrected on the axis as the doublets. It is usually not a good idea to attempt to correct the Petzval curvature of the relay lenses until they start to introduce a contribution which is one half, or at most equal to, that of the eyepiece.

12.4 SECONDARY COLOR IN A RELAY SYSTEM

12.4.1 In paragraph 6.10.8.3 it was indicated that telescope lenses made out of ordinary glass have an amount of secondary color given by the expression

$$TAch_{F-D} = \frac{y_1}{2200} .$$

12.4.2 The relay lens in the sample problem in Figure 12.1 would have secondary color given by the equation

$$TAch_{F-D} = \frac{2y_1}{2200} = \frac{u_o D}{2200} . \qquad (1)$$

12.4.3 Equation (1) shows that the secondary blur at the focal plane of the eyepiece increases as u_o or D is increased. In the sample problem, if $u_o = 0.14$ and $D = 70$ cm., the radius of the secondary blur $TAch_{F-D}$ is 0.005 cm. With the 10X eyepiece, this would subtend an angle of 0.002 radians. This is 6.6 minutes, which is definitely noticeable but usually tolerated. Any more than this is objectionable.

12.4.4 Secondary color is usually a serious problem in relay systems. If the distance D must be maintained, then the secondary color can most easily be reduced by making u_o smaller. A value of $u_o = 0.14$ means that the exit pupil diameter will be 7 mm. This is desirable for maximum light transmission, but it could be reduced to 2 mm without impairing the observer's resolution. This would then cut down the secondary color to 2.2 minutes.

12.4.5 The secondary color can be reduced by separating the two doublets. As long as there is parallel light between the two lenses, the space between the lenses can be considered free space. If this distance is d, the secondary color is given by the equation

$$TAch_{F-D} = \frac{u_o (D-d)}{2200} . \qquad (2)$$

12.4.6 As d is made larger, the focal lengths of the lenses are reduced. They therefore introduce more field curvature. It is a fairly general rule that any step taken to reduce secondary color without sacrificing clear aperture will result in more field curvature.

12.5 FURTHER DETAILS ON DESIGN OF DOUBLETS AS RELAY LENSES

For a unit power relay system, there are advantages in using two identical doublets with parallel light between them. Since the doublets are usually air-spaced, this means there are eight glass air surfaces. In principle it is possible to combine the positive elements of the doublets into a single lens surrounded by two negative lenses. One could also combine the two negative lenses and surround the combination with two positive elements. One can see what would be involved in doing this by considering the solutions shown in Table 11.3. In these solutions, the curvature facing the parallel light c_1 is about 0.16. In order to make the positive element in the combined doublets a single lens, it would be necessary to bend these solutions until $c_1 = 0$. Take, for example, Case No. 10 in Table 11.3. If the lens should be bent to make $c_1 = 0$, the remaining curvatures would be $c_2 = -0.4798$, $c_3 = -0.4694$, and $c_4 = -0.2341$. The doublet would no longer be corrected for spherical aberration or coma. The spherical aberration, at least to the third order, could be corrected by adjusting the curvature differences between c_2 and c_3. The coma would, however, be far from corrected. By facing this with an identical doublet, the two plane surfaces could be contacted. This means the positive lens could be made into a single equiconvex lens. The spherical aberration of the doublets would add, but the coma would subtract, to zero. This argument shows that a triplet relay lens is possible, but it also shows that it will probably have considerably more higher order spherical aberration than the two doublets. The use of a triplet of this type could be recommended only when the zonal spherical aberration is tolerable.

12.6 DOUBLE-RELAY SYSTEMS

12.6.1 Often in relay systems one is limited in the diameter of lens which may be used as, for example, when the lens must be confined to a given diameter. In the single-relay example described in 12.3.3, the relay was allowed to have a diameter of 10 cm. Suppose there was a limitation to a 5-cm. diameter for all lenses. In order to maintain $u_o = 0.14$, it would be necessary to use a double-lens relay system as illustrated in Figure 12.2.

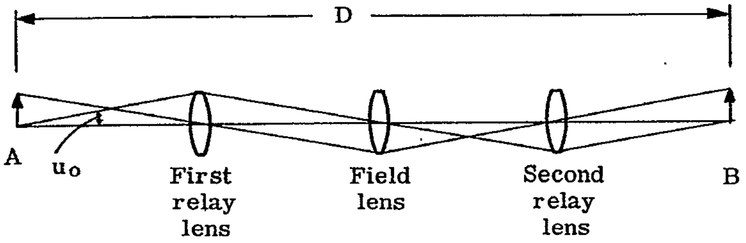


Figure 12.2 - A double-relay system.

12.6.2 In the double-relay lens system, the first and second relay lenses must have just one-half the focal length of the single-relay lens. They will, therefore, each add twice the Petzval contribution.

12.6.3 The two relay lenses will not introduce any more secondary color. Each relay transfers by the distance D/2, so that each has half the secondary color; but they add, so the total comes out the same. Equation (1) still applies for a double relay system.

12.6.4 Note that in Figure 12.2 an extra lens has been added in the intermediate focal plane. This is called a field lens. Its function is to image the chief ray passing through the center of the first relay lens at the center of the second relay lens. This field lens has a focal length equal to D/8. It will also introduce negative field curvature equal to 8/Dn.

12.6.5 One can see then that doubling up the relay system in order to reduce the diameter of the lenses has introduced Petzval field curvature. The relay lenses introduce four times as much field curvature, and the field lens adds as much as one of the relays. The double-relay lens, therefore, introduces six times as much field curvature as the single-relay system of equal length and numerical aperture. The secondary color is not changed.

12.6.6 Reducing the field of view. One can argue that there is little to be gained in reducing the field of view. The eyepiece is designed for, and capable of, viewing an object height of 1.32 cm. It is true that the relay lenses are going to make the image at the end of the field more blurred, but to introduce a stop at the field lens would merely mean a slightly smaller field lens. The savings in cost would be negligible. If one decides to use doublet relay lenses, nothing is lost in using the full field of the eyepiece. It is better to see the wide field, even if it is blurred, than to stop it down. As a rule, the field of a visual instrument should not be reduced if the extra field can be obtained with no increase in cost and size of instrument.

12.6.7 Relay lenses corrected for field curvature. If the problem demands improved quality at the edge of the field, it is necessary to abandon the doublet relay lenses and use a lens with reduced field curvature. Triplets as described in Section 10 may be used if one does not try to increase n_e beyond 4. If this is not enough, a double Gauss lens is recommended.

12.6.8 Field lenses corrected for field curvature. It is possible to introduce compound field lenses, such as triplets, to reduce the Petzval curvature of the field lens. If the field lens is located between the two unit power relay systems, it must be symmetrical. It will have to perform for a finite conjugate, and the region of solution will be quite different from the lenses described in Section 10. One can think of the lens as basically two inverted telephoto objectives with parallel light between them. The lens can be roughed out by first designing one side. One should avoid placing a lens surface directly in the intermediate focal plane, for it will eventually collect dust.

12.7 SUMMARY

12.7.1 Relay systems are inherently limited by problems of field curvature and secondary color. One should always try to use as few relays as possible, until glass weight and cost become a problem.

12.7.2 If it is necessary to use more than one relay, the same rules apply. The relay and field lenses should be kept as large as practicable. The size of the relay or field lenses should never be reduced needlessly.

12.7.3 The secondary color depends on the over-all distance of relay and the value of u_0. If the secondary color becomes serious, it is necessary either to accept it, reduce u_0, or resort to special lens materials.

12.7.4 Whenever attempts are made to correct the field curvature, more secondary color is introduced. See Figure 6.21.

12.7.5 Doublet relay lenses are usually prefered to other more complicated types. They represent a good compromise between simplicity, cost, number of surfaces, and reasonable image quality. It is possible rapidly to reach a point of diminishing returns in trying to reduce field curvature: the result will be a design with increased secondary color and chromatic variation of aberrations.

13 MIRROR AND PRISM SYSTEMS

13.1 INTRODUCTION

13.1.1 Uses of mirrors and prisms. Mirrors and prisms are widely used in optical systems. Among the principal uses are the following:

(1) To bend light around corners.

(2) To fold an optical system into a smaller space.

(3) To provide proper image orientation.

(4) To combine or split optical beams with partial reflecting surfaces.

(5) To disperse light, as in refractometers and spectrographic equipment.

13.1.2 Design application. The principles discussed in this section are intended to develop an understanding of concepts, and to provide computational tools for use in designing optical systems for all the above applications with the exception of spectrographic equipment. Thus, since dispersion is not one of our primary aims, the problem can best be approached by the study of reflection.

13.2 REFLECTION

13.2.1 Reflection from a single surface.

13.2.1.1 The first problem involved in the study of reflecting surfaces is illustrated in Figure 13.1. An object point P is given. A mirror reflects the incident rays of light from P in a new direction so that the reflected rays appear to emerge from an image P'. The actual reflection problem might involve a number of possible variations from a design standpoint. For example, the problem might be to orient the mirror to send the reflected light in a given direction. This might then raise the question of image orientation at P'.

13.2.1.2 The simpler problems of this nature can be readily solved by elementary concepts known to most technical people. The discussion below is designed to provide the tools to handle more complex problems.

13.2.2 Multiple reflection.

13.2.2.1 Equations 2-(3) and 2-(4) provide a vector form for the law of refraction and the law of reflection. The same equations can be used to treat reflection problems by assuming that, $-n_1 = n_o = 1$. From equation 2-(4),

$$\Gamma = -\cos I - \sqrt{\cos^2 I}\ , \text{ or}$$

$$\Gamma = -2\cos I. \qquad (1)$$

Cos I is given by the dot product, $\vec{S}\cdot\vec{M}$; therefore

$$\Gamma_i = -2(\vec{S}_{i-1}\cdot\vec{M}_i) = -2\rho_i\ . \qquad (2)$$

Equation 2-(3) and Equation (2) above make it possible to handle reflection problems for any number of surfaces. For example, assume a system of mirrors as in Figure 13.2, with rays reflected as illustrated. If $\vec{S}$ is a unit vector along any ray, thereby indicating its direction, it is possible to write the following equations.

$$\vec{S}_1 = \vec{S}_o + \Gamma_1\vec{M}_1, \qquad (3a)$$

$$\rho_1 = \vec{S}_o\cdot\vec{M}_1, \qquad (3b)$$

and

$$\vec{S}_2 = \vec{S}_1 + \Gamma_2\vec{M}_2, \qquad (4a)$$

$$\rho_2 = \vec{S}_1\cdot\vec{M}_2 = \vec{S}_o\cdot\vec{M}_2 + \Gamma_1\vec{M}_1\cdot\vec{M}_2, \qquad (4b)$$

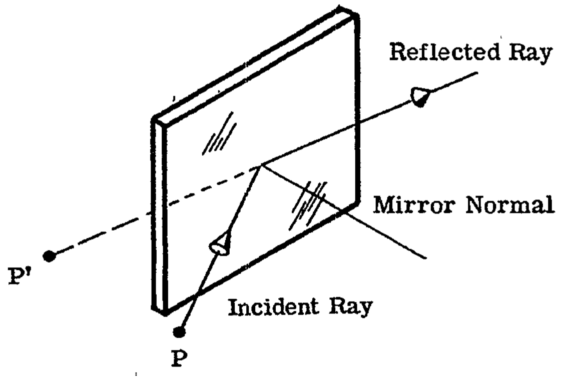


Figure 13.1-Reflection from a single surface mirror.

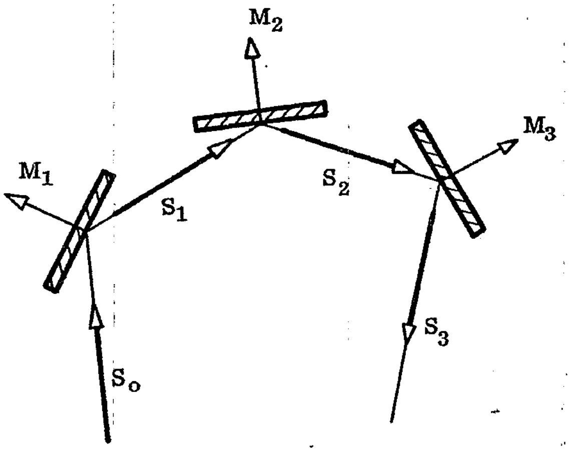


Figure 13.2-Reflection from multiple mirrors.

and

$$\vec{S}_3 = \vec{S}_2 + \Gamma_3 \vec{M}_3, \tag{5a}$$

$$\rho_3 = \vec{S}_2 \cdot \vec{M}_3 = \vec{S}_o \cdot \vec{M}_3 + \Gamma_1 \vec{M}_1 \cdot \vec{M}_3 + \Gamma_2 \vec{M}_2 \cdot \vec{M}_3, \tag{5b}$$

from which one can readily see the pattern that follows as more surfaces are added.

13.2.2.2 Let us examine an example of a problem involving a single reflection. Suppose it is desired to have a ray of light pass along the Z axis and reflect from a mirror in the XY plane at an angle of 45° to the X axis as in Figure 13.3. What are the coordinates of the normal to the mirror? By writing the incoming and outgoing vectors in component form, we have

$$\vec{S}_o = \vec{k},$$

and

$\vec{S}_1 = \frac{1}{\sqrt{2}} \vec{i} + \frac{1}{\sqrt{2}} \vec{j}$, where $\vec{i}$, $\vec{j}$ and $\vec{k}$ are unit vectors along the X, Y and Z axes, respectively. The unit vector for the mirror normal may then be written as

$$\vec{M} = M_x \vec{i} + M_y \vec{j} + M_z \vec{k}.$$

Therefore

$$\rho_1 = \vec{S}_o \cdot \vec{M}_1 = M_z$$

and

$$\Gamma_1 = -2 M_z.$$

Then, from equation (3a),

$$\frac{1}{\sqrt{2}} \vec{i} + \frac{1}{\sqrt{2}} \vec{j} = \vec{k} - 2M_z \vec{M}.$$

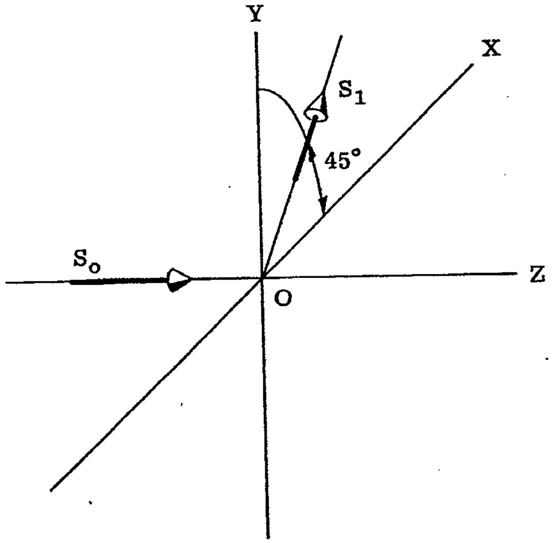


Figure 13.3-A single reflection problem.

It follows that

$$\vec{M} = -\frac{1}{2\sqrt{2}\,M_z}\vec{i} - \frac{1}{2\sqrt{2}\,M_z}\vec{j} + \frac{1}{2\,M_z}\vec{k}.$$

Since $\vec{M}$ is a unit vector, the sum of the squares of its components is equal to one. Therefore,

$$\left(\frac{1}{2\sqrt{2}\,M_z}\right)^2 + \left(\frac{1}{2\sqrt{2}\,M_z}\right)^2 + \left(\frac{1}{2\,M_z}\right)^2 = 1$$

since $\vec{i}$, $\vec{j}$ and $\vec{k}$ are also unit vectors.

Solving for M_z,

$$M_z^2 = \frac{1}{8} + \frac{1}{8} + \frac{1}{4},$$

$$M_z = \frac{1}{\sqrt{2}}.$$

Finally,

$$\vec{M} = -\frac{1}{2}\vec{i} - \frac{1}{2}\vec{j} + \frac{1}{\sqrt{2}}\vec{k}.$$

From this we can see that

$$M_x = -\frac{1}{2},\quad M_y = -\frac{1}{2} \quad \text{and} \quad M_z = \frac{1}{\sqrt{2}}$$

13.2.2.3 Consider the above solution. $\vec{M}$ is the vector for the mirror normal, but what is the significance of describing it thusly? We will find it very convenient to be able to describe the equations of a plane in terms of the components of a unit vector normal to the plane. The equation of a plane may be written as

$$Ax + By + Cz + D = 0. \tag{6}$$

Taking the numerical value of D as negative, if P is the distance from the origin to the plane along the normal,

$$P = \frac{-D}{\sqrt{A^2 + B^2 + C^2}} = \frac{-D}{F},$$

where

$$F = \sqrt{A^2 + B^2 + C^2} . \tag{7}$$

The components of P on the X, Y, Z axes are,

$$P_x = -\frac{DA}{F^2}, \tag{8a}$$

$$P_y = -\frac{DB}{F^2}, \tag{8b}$$

and

$$P_z = \frac{DC}{F^2} . \tag{8c}$$

The coordinates of the unit vector along P are, therefore,

$$M_x = -\frac{A}{F}, \tag{9a}$$

$$M_y = \frac{B}{F}, \tag{9b}$$

and

$$M_z = \frac{C}{F}. \tag{9c}$$

These equations enable us to visualize the spatial position of the mirror discussed above. If $P = 1$, then $F = -D$ and the intercepts of the mirror on the X, Y, Z axes are equal to $\frac{1}{M_x}$, $\frac{1}{M_y}$ and $\frac{1}{M_z}$ because,

$$-\frac{D}{A} = \frac{1}{M_x}, \quad -\frac{D}{B} = \frac{1}{M_y}, \quad \text{and} \; -\frac{D}{C} = \frac{1}{M_z} .$$

In the above example, then, the intercepts of the plane of the mirror are,

$$\frac{1}{M_x} = -2, \quad \frac{1}{M_y} = -2, \quad \text{and} \quad \frac{1}{M_z} = \sqrt{2} .$$

A plane mirror located with these intercepts will be parallel to the mirror specified in the problem, and at a distance $P = 1$ from it as shown in Figure 13.4. (The intercepts of the desired plane, of course, are 0, 0, 0.) The components of the mirror normal vector for the mirror at the origin will be equal to the components of mirror normal vector for the mirror at P since the mirrors describe two parallel planes.

13.3 LOCATION OF THE IMAGE

13.3.1 The plane of incidence. One of the conditions of the law of reflection is that the incident ray, a normal to the surface at the point of incidence, and the reflected ray all lie in a single plane. It is possible therefore to draw the plane containing the incident ray, the normal to the surface, and the reflected ray. This is illustrated in Figure 13.5. The plane containing this ray is called the plane of incidence.

13.3.2 Image location.

13.3.2.1 The next problem of interest is the following. If point P represents an object point, where will its image be located? In order to locate an image it is necessary to take at least two rays from the object point and reflect them from the mirror. These are indicated by R_1 and R_2 in Figure 13.5. One can readily determine that the second ray, when extended back, intersects the first ray at P'. P' is therefore the image of P; it is located on the line from P perpendicular to the mirror and lies behind the mirror the same distance that P is in front of the mirror.

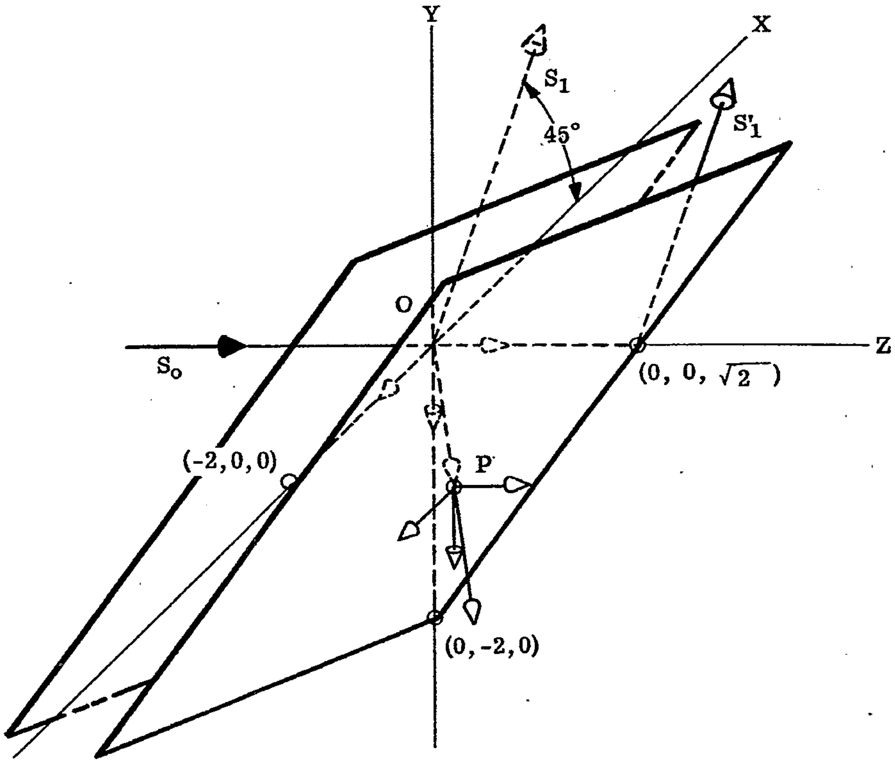


Figure 13.4-Solution to problem of Figure 13.3.

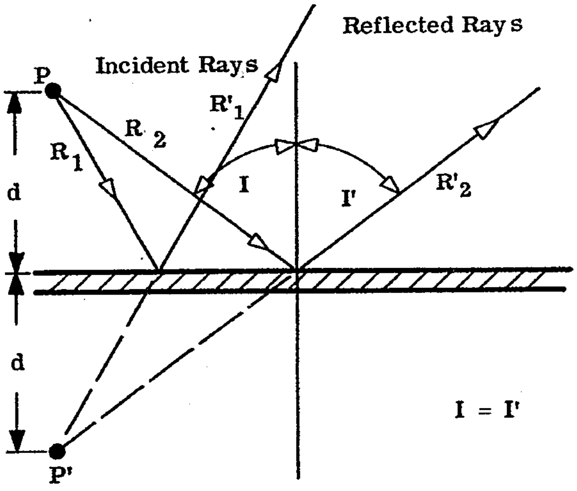


Figure 13.5-Plane of incidence.

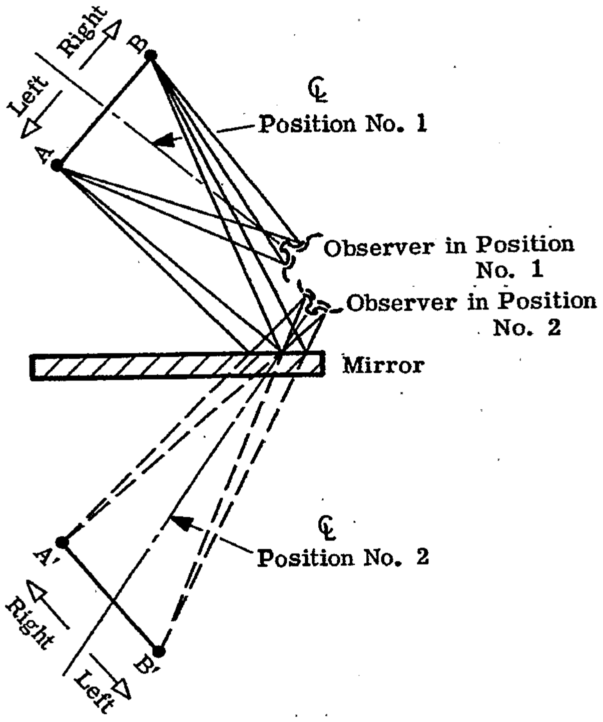


Figure 13.6-Observer, image, and object positions.

13.3.2.2 This means that the image of a point P in a mirror may be located immediately by drawing a line from P perpendicular to the mirror. If the distance along this line from P to the mirror is d, then the image P' will be located on this same line in back of the mirror at a distance d from the mirror. Alternately the image of a point P in a mirror may be found by rotating the object point around the axis formed by the intersection of the plane of the mirror and the plane of incidence.

13.4 ORIENTATION OF THE IMAGE

13.4.1 Single mirror imagery. Suppose we look at the image of two points A and B. See Figure 13.6. The images A' and B' are located readily by drawing normals through the mirror and laying off equal distances. Now suppose that an observer looks at the AB from position 1, shown. To the observer B lies to the right of A. Now if the observer wishes to see the image he must turn around and look into the mirror as in position 2. Then B' appears to lie to the left of A'. This means the mirror image appears to be "left handed". An object imaged by a single mirror always appears "left handed". One source of confusion in this field stems from the fact that one may not always look at the object from the same side. Figure 13.6 shows that A' and B' are actually in the same spatial orientation as A and B. It is because the observer has to change his point of view that makes the image appear left handed.

13.4.2 Mathematical formulae for locating the image of a point P in a mirror.*

13.4.2.1 It is possible to readily compute the image position of an object point P as reflected in a mirror. Referring to Figure 13.5, one may write the expression for a plane parallel to the mirror passing through P. The equation is

$$A(x_1) + B(y_1) + C(z_1) + D_1 = 0 \ . \qquad (10)$$

This represents a plane through P which is located at coordinates x_1, y_1, z_1. The equation for the mirror is

$$A(x) + B(y) + C(z) + D = 0 \qquad (11)$$

The perpendicular distance between the two planes is therefore

$$d = \frac{D-D_1}{F} = \frac{A(x_1) + B(y_1) + C(z_1) + D}{F} \ . \qquad (12)$$

The image will lie at a distance d on the other side of the mirror from the point P on the normal to the mirror. Equation (9) gives the components for the unit vector perpendicular to the mirror, so if these are multiplied by 2d, one obtains the differences in the position coordinates for the object and image. The position coordinates of the image P' (x_1', y_1', z_1') are then given by

$$x_1' = x_1 - 2d\,\frac{A}{F} \ , \qquad (13)$$

$$y_1' = y_1 - 2d\,\frac{B}{F} \ , \qquad (14)$$

and

$$z_1' = z_1 - 2d\,\frac{C}{F} \ . \qquad (15)$$

By inserting the value of d from Equation (12), it is possible to compute x_1', y_1' and z_1'.

13.4.2.2 It is convenient to use matrix notation for Equations (13), (14), and (15). These equations may be written in matrix form as follows,

$$\begin{bmatrix} 1 \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2AD/F^2 & 1-2A^2/F^2 & -2BA/F^2 & -2AC/F^2 \\ -2BD/F^2 & -2AB/F^2 & 1-2B^2/F^2 & -2BC/F^2 \\ -2CD/F^2 & -2AC/F^2 & -2BC/F^2 & 1-2C^2/F^2 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \\ z \end{bmatrix} \qquad (16)$$

---

*J. S. Beggs, J. Opt. Soc. Am. 50, 388 (1960).

In abbreviated form, then, one can say,

$$\left[P\right] = \left[M\right] \left[P'\right], \tag{17}$$

in which P represents the column matrix, $\begin{bmatrix} 1 \\ x' \\ y' \\ z' \end{bmatrix}$,

and P' represents the column matrix, $\begin{bmatrix} 1 \\ x \\ y \\ z \end{bmatrix}$.

The M matrix is the large matrix made up of the constants of the mirror. Now if there are several mirrors involved, the image P' will be transformed to another image P'' and P'' to P''' etc. It follows that

$$\left[P'\right] = \left[M_1\right] \left[P\right] \text{ and } \left[P''\right] = \left[M_2\right] \left[P'\right] \text{ etc.,}$$

$$\therefore \left[P^n\right] = \left[M_n\right] \cdots\cdot \left[M_2\right] \left[M_1\right] \left[P\right]. \tag{18}$$

13.4.3 The vector ray tracing equation in matrix form.

13.4.3.1 Equations (3a) and 3b) may also be written in matrix form. First combine (3a) and (3b),

$$\vec{S_1} = \vec{S_0} - 2 (\vec{S_0} \cdot \vec{M_1}) \vec{M_1}.$$

In component form this equatiom may be written

$$S_{1x} = S_{0x} - 2 M_x (S_{0x} M_x + S_{0y} M_y + S_{0z} M_z),$$

$$S_{1y} = S_{0y} - 2 M_y (S_{0x} M_x + S_{0y} M_y + S_{0z} M_z);$$

and

$$S_{1z} = S_{0z} - 2 M_z (S_{0x} M_x + S_{0y} M_y + S_{0z} M_z).$$

These equations may also be reduced to matrix form,

$$\begin{bmatrix} S_{1x} \\ S_{1y} \\ S_{1z} \end{bmatrix} = \begin{bmatrix} (1-2M_x^2) & -2M_x M_y & -2M_x M_z \\ -2M_x M_y & (1-2M_y^2) & -2M_y M_z \\ -2M_x M_z & -2M_y M_z & (1-2M_z^2) \end{bmatrix} \begin{bmatrix} S_{0x} \\ S_{0y} \\ S_{0z} \end{bmatrix}. \tag{19}$$

By substituting Equations (9a), (9b) and (9c) into Equation (16), we see that this new matrix is a minor of the M matrix. Let us call this the R matrix. For several reflections, then, it is possible to write

$$\left[S_n\right] = \left[R_n\right] \left[R_{n-1}\right] \cdots\cdot \left[R_1\right] \left[S_0\right]. \tag{20}$$

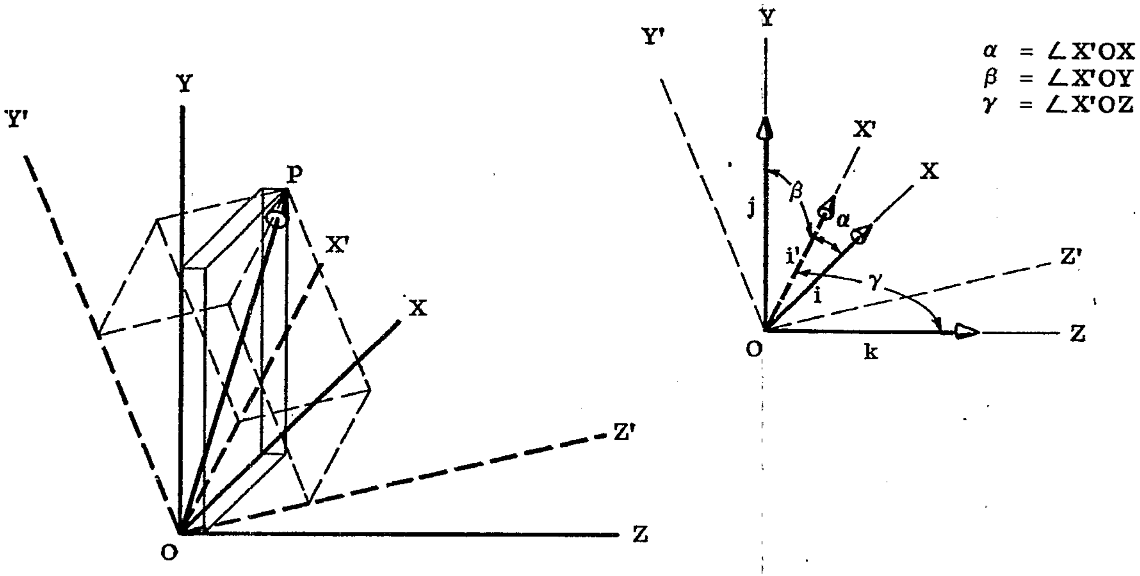
13.4.3.2 The matrix notation is conceptionally convenient because the matrix equation (19) represents a rotation of coordinate axes. To illustrate, consider the rectangular coordinate axes X, Y, Z and their respective unit vectors i, j, k and the rotated coordinate axes X', Y', Z' and their unit vectors i', j', k'. The vector $\overline{OP}$ shown in Figure 13.7 may be written in component form for either system of coordinates as,

$$\vec{OP} = xi + yj + zk = x'i' + y'j' + z'k'.$$

Performing scalar multiplication by i yields

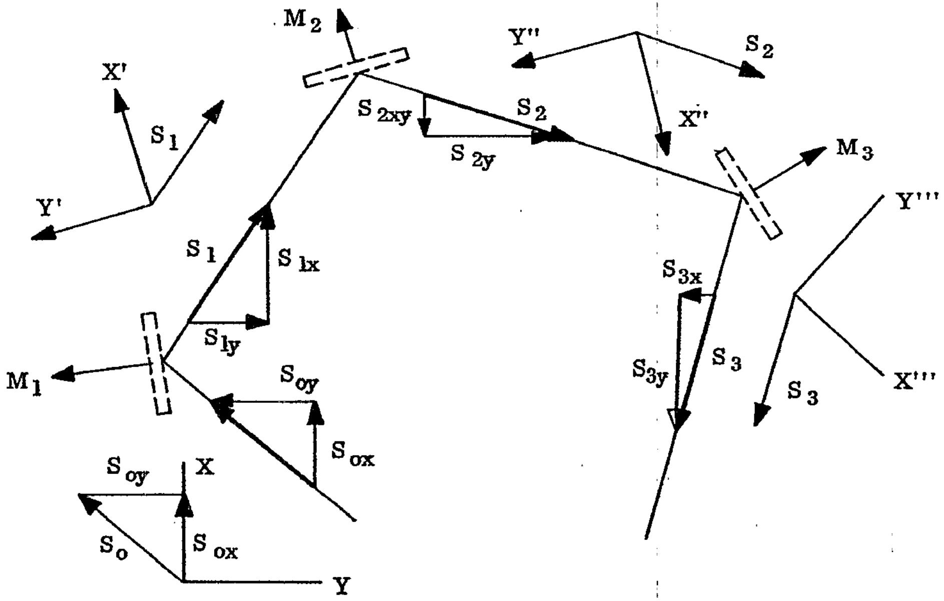
$$x (i \cdot i) + y (i \cdot j) + z (i \cdot k) = x' (i \cdot i') + y' (i \cdot j') + z' (i \cdot k'). \tag{21}$$

If we let l_1, m_1, and n_1, be the direction cosines for the X' axis in the XYZ coordinate system, where the



Position coordinates of the point P are P (x, y, z) in the XYZ system, and P (x', y', z') in the X'Y'Z' system

Figure 13.7-Rotation of the coordinate axes.



Ray Unit Vector	Components in XYZ System	In Mirror Image System	
$\vec{S_o}$	S_{ox} and S_{oy}	$S_{ox} = S_{ox}$	$S_{oy} = S_{oy}$
$\vec{S_1}$	S_{1x} and S_{1y}	$S_{1x'} = S_{ox}$	$S_{1y'} = S_{oy}$
$\vec{S_2}$	S_{2x} and S_{2y}	$S_{2x''} = S_{ox}$	$S_{2y''} = S_{oy}$
$\vec{S_3}$	S_{3x} and S_{3y}	$S_{3x'''} = S_{ox}$	$S_{3y'''} = S_{oy}$

Figure 13.8 - Diagram showing how the mirrors cause rotation of the coordinate system.

direction angles are α_1, β_1 and γ_1, respectively, then, the dot product,

$$i'.i = (i' \cos \alpha_1)\ i = \cos \alpha_1,$$

since i' and i are unit vectors, and similarly;

$$i' \cdot j = \cos \beta_1,$$

and

$$i' \cdot k = \cos \gamma_1.$$

We may then let

$$i' \cdot i = l_1, \quad i' \cdot j = m_1 \quad \text{and} \quad i' \cdot k = n_1,$$

and, similarly,

$$j' \cdot i = l_2, \quad j' \cdot j = m_2, \quad j' \cdot k = n_2,$$

$$k' \cdot i = l_3 \quad k' \cdot j = m_3, \quad k' \cdot k = n_3,$$

where l_2, m_2 and n_2 are direction cosines of the Y'-axis and l_3, m_3 and n_3 are direction cosine of the Z'-axis respectively, in the XYZ coordinate system. We may now rewrite Equation (21):

$$x = x' l_1 + y' l_2 + z' l_3, \tag{22}$$

$$y = x' m_1 + y' m_2 + z' m_3, \tag{23}$$

and

$$z = x' n_1 + y' n_2 + z' n_3. \tag{24}$$

These three equations may be written in the matrix form,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{bmatrix} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}. \tag{25}$$

By similar reasoning, it can be shown that

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \tag{26}$$

13.4.4 Interpretation of the vector matrix.

13.4.4.1 Note that the above equations are exactly similar to Equation (19) which, therefore, can be thought of in the following way. The object ray has the direction cosines S_{0x}, S_{0y}, S_{0z} with respect to the $x_1 y_1 z_1$ coordinate axis. After reflection it has the direction cosines S_{1x}, S_{1y}, S_{1z} in the same coordinate system. See Figure 13.8. Another way to look at it is that reflection has caused a rotation of the coordinate system. The direction cosines of the new coordinate system with respect to the old are given by the terms in the reflection matrix R. This is a very convenient concept because it gives directly the rotation between the object and its image. There is a great deal known about rotation matrices. For example if the determinant of the matrix is -1, it means the image coordinate system is left-handed. One can check the determinant in the R matrix in Equation (19) and see that it is -1. This follows from the condition that M is a unit vector, and

$$M_x^2 + M_y^2 + M_z^2 = 1.$$

13.4.4.2 By Equation (20) it is evident that if there are an even number of reflections the determinant of the total reflection matrix is +1 while if there are an odd number of reflections the determinant of the matrix is -1.

This is stated in optics in the following way.

(1) An image seen by an even number of reflections is right-handed.

(2) An image seen by an odd number of reflections is left-handed.

A left-handed image of a readable page of print is not readable. A right-handed image of a readable page of print is readable. It may be turned at an odd angle, even upside down, but the observer can read it by standing on his head. A left-handed image is always backwards regardless of the orientation of the image. In Figure 13.9 the letter R is shown as left handed and right handed. The right-handed image may be made to appear normal by turning the paper around. The paper cannot be rotated into a position which will make the left-handed image readable.

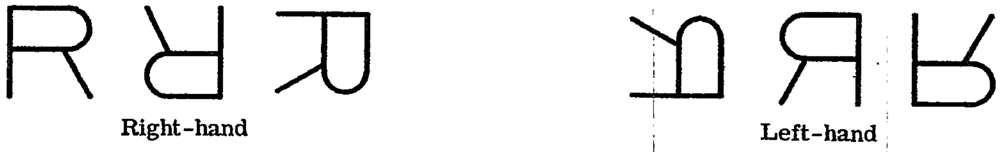


Figure 13.9-The right and left-hand image.

13.5 THE IMAGE SPHERE

13.5.1 The external observer concept.

13.5.1.1 Some people find it helpful in understanding the imagery of a single mirror to make use of the image sphere shown in Figure 13.10.

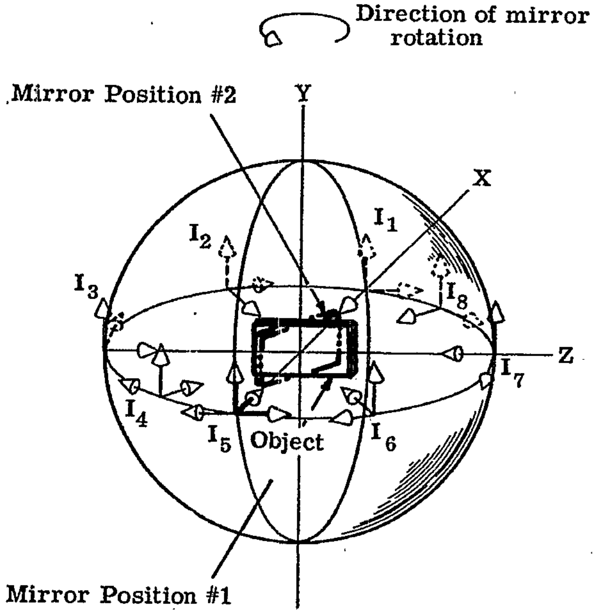


Figure 13.10-Image position and orientation in the Y-plane.

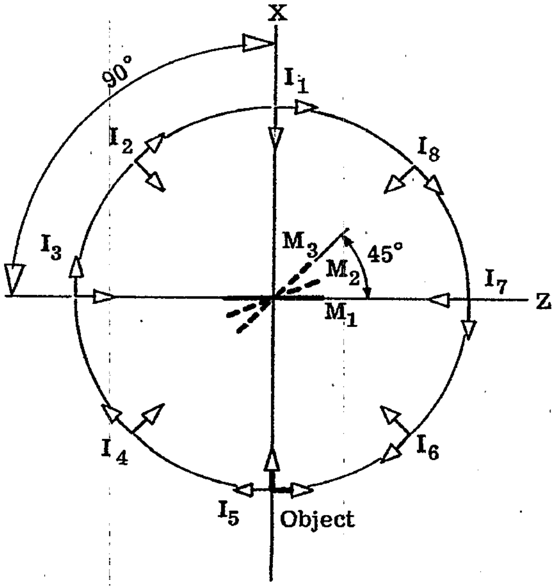


Figure 13.11-The Y-plane mirror rotation.

Suppose that an object represented by a small coordinate axis is located at 0° azimuth and 0° elevation as shown in Figure 13.10. Now imagine placing a mirror in the center of the sphere. By rotating this mirror the vertical images may be made to appear at any position of the surface of the sphere. For example, consider that we are looking directly down on the XZ plane of the sphere. Figure 13.11 shows this view. A plane mirror mounted in the center with its plane vertical, and facing the object as in position M_1, will produce a virtual image at I_1 as shown. This is very easily demonstrated by placing a small pocket mirror in position M_1 on Figure 13.11.

13.5.1.2 Now, as the mirror is rotated about the vertical axis (the Y axis) to position M_2, the image shifts to I_2, and similarly with M_3 and I_3 and so on until the image swings completely around in the horizontal plane. If you are using a pocket mirror, you will note that the image position and orientation coincides exactly with that drawn, regardless of the observer's position. Of course, the observer must place himself so that he can see the image to confirm this. The significance is that the image does have spatial position and orientation whether observed or not, and that this is related only to the object and mirror relationship.

13.5.1.3 Consider now, the image position shift in relation to the mirror. As the plane of the mirror was rotated through an angle of 45° from M_1 to M_3 the image position shifted through an angle of 90°.

13.5.1.4 Vertical relations are similar. If the mirror placed initially in the position shown in Figure 13.12, and then rotated about the horizontal axis (Z axis), the image will assume the positions and orientations shown. Figure 13.13 shows a projection of the XY plane. Experiments with a plane mirror will again confirm the accuracy of the illustrations, if the observer remembers that the Z axis is pointing "up" from the paper.

13.5.1.5 To make full use of this concept, Figure 13.14 illustrates the position and orientation of the image for compound angles. In each case, the mirror has been tipped 22 1/2° from vertical and rotated 22 1/2° from the Z axis in the XZ plane.

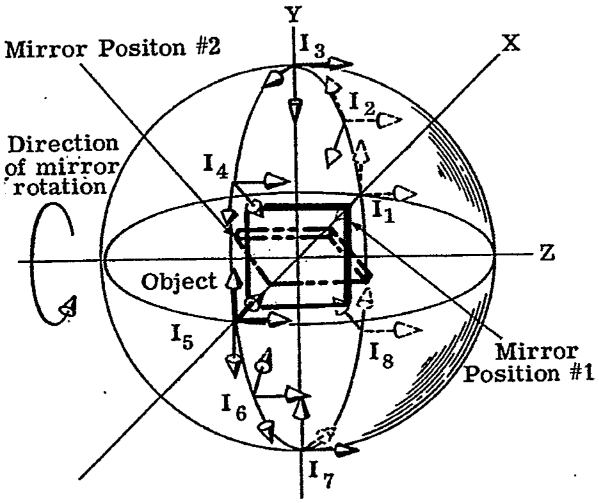


Figure 13.12-Image position and orientation in the Z-plane.

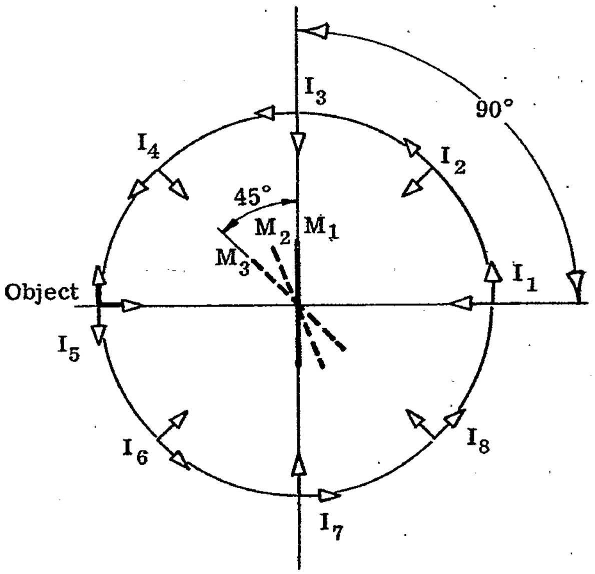


Figure 13.13-Projection of the XY-plane.

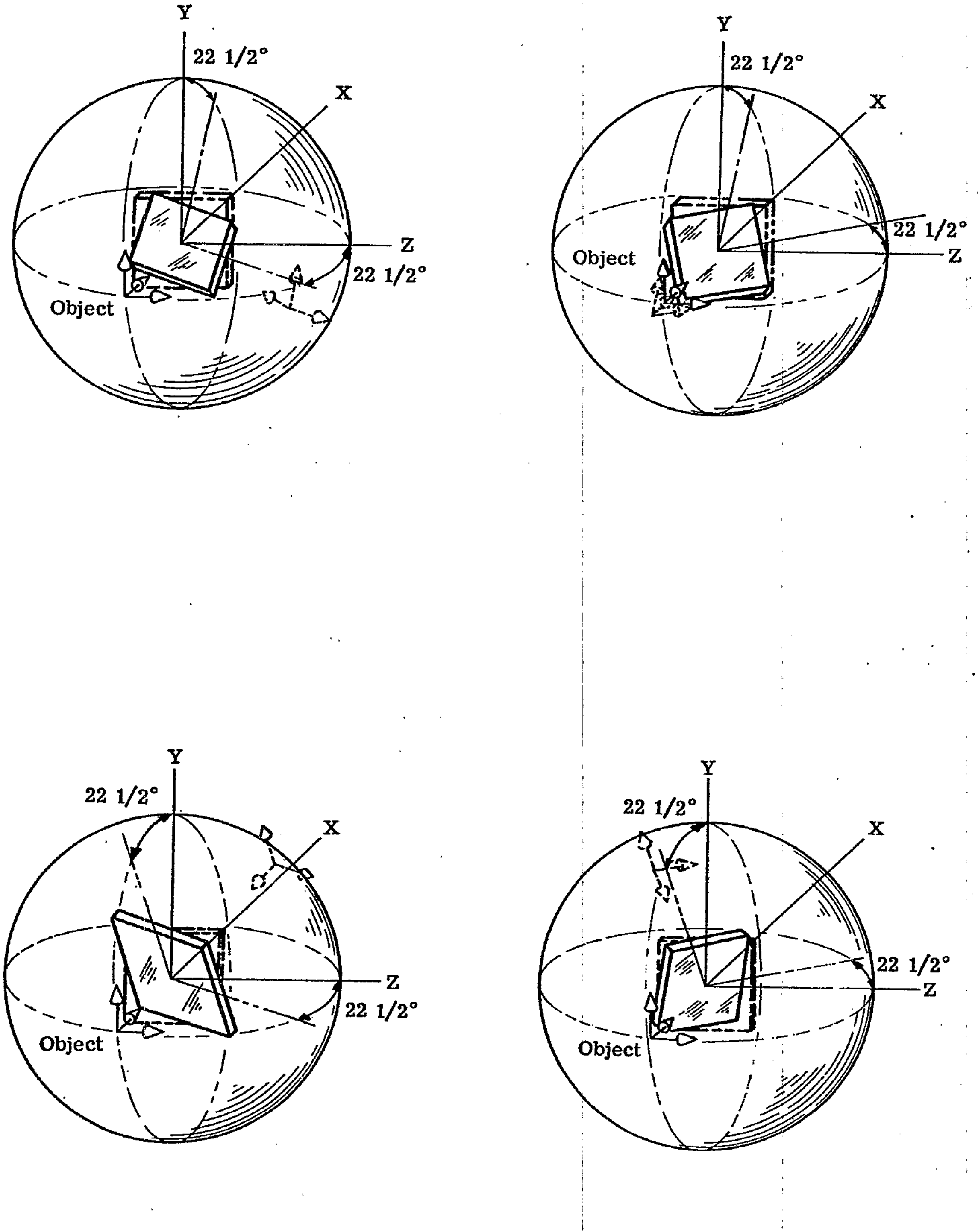


Figure 13.14-Image position and orientation for compound angles.

13.5.2 The internal observer concept.

13.5.2.1 It may be more convenient to visualize this in the following way. Imagine you are in a large sphere at the center. Assume that the X axis is due North and South. The Z axis is the East and West and the Y axis is straight up and down. See Figure 13.15(d). Along side of you is a projector, projecting an image on the inside of the sphere due south on the horizon. By placing a mirror in front of the projector the virtual images may be projected to any position on the sphere. See Figure 13.15.

13.5.2.2 First consider the case where the mirror reflects the light just east or west of due south at 0° elevation. It will not be possible to project it exactly where the original projected image is for then the plane of the mirror would be exactly parallel to the mirror, but it would be possible to reflect some light a few degrees to the east or west. The projected image would then appear as shown in Figure 13.15. As the mirror is rotated and the images are always located at an equal angular position around the object, they appear to be rotated. When the images are located in the horizontal plane it appears left handed but erect. (The y' axis is in the same direction as the y axis). When the image appears in the vertical plane it appears left handed but upside down. As the image is rotated through 90° its orientation turns 180°. Intermediate positions are linearly connected.

13.5.2.3 This concept enables us to predict the orientation for the position of the image at any position on the sphere. To do this one uses the following reasoning. Suppose one wishes to project an image on the inside of the sphere at a point with an azimuth angle of 45° and an elevation angle of 30°. If one images a cone with its apex at the center and its axis along the X axis, it will pierce the sphere at a circle. This circle is the one shown in Figure 13.12. This circle defines a plane. Images on this circle rotate twice as fast as the angle θ between the Y plane and a line drawn perpendicular from the X axis to the image point P. Therefore if θ can be calculated, the rotation of the image is known. Figure 13.15 illustrates the above case.

$$\tan\theta = \frac{\tan\phi}{\sin\omega}$$

where ω = azimuth angle and ϕ = elevation angle.

In the above cited example $\omega = 45°$ and $\phi = 30°$.

$$\tan\theta = \frac{.5774}{.7071} = 0.81657$$

$$\theta = 39.2°.$$

The image will therefore have been rotated by 2θ or 78.4°.

13.6 REFLECTION FROM TWO MIRRORS

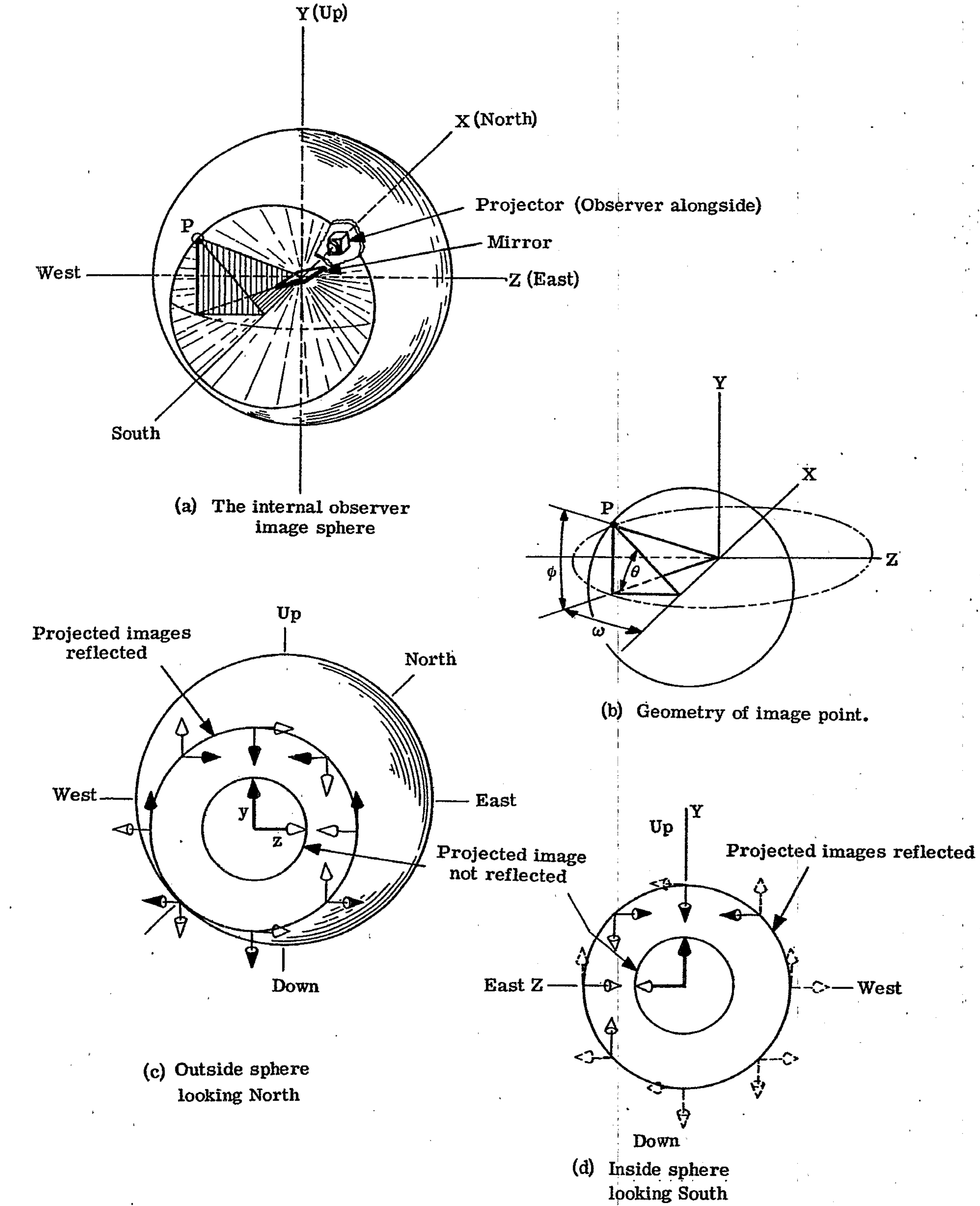
13.6.1 Location of the image.

13.6.1.1 In the case of reflection from a single mirror, the image may always be located by projecting a line from the object perpendicular to the mirror and locating the image on the extension of this perpendicular at an equal distance behind the mirror as in paragraph 13.3. For the double mirror system the image is located in a plane perpendicular to the intersecting edges.

13.6.1.2 Figure 13.16 illustrates a special case of this. In the illustration, the two mirrors are perpendicular. The image points P' and P" have been located by first constructing the perpendicular from P to mirror #1 and locating P' as above. Then, using P' as the object point for mirror #2, the same procedure was used to locate P". It is, therefore, evident that the perpendiculars PP' and P'P" lie in the plane PP'P". Now, since mirrors #1 and #2 are perpendicular to PP' and P'P" respectively, the intersection of their planes, LL' is perpendicular to the plane PP'P". From the illustration one can see that the image P" formed by the second reflector lies on the line PP" and that this line intersects LL' and is perpendicular to it. In a more general case where the mirrors are not perpendicular, the plane PP'P" will still be perpendicular to LL' but the line PP" will not intersect LL'.

13.6.2 Axiom for locating the image. The location of the image in a perpendicular double-mirror system may be found by projecting a line from the object point through, and perpenducular to the line of intersection of the mirror surfaces. The image may lay on the extended perpendicular an equal distance behind the line of intersection and will be right handed since there are two reflections.

13.6.3 Invariant position of the image. Since the image in a double mirror lies in a plane normal to the intersecting edge of the two mirrors, the positioning of the image depends on the position in space of the intersecting edge. If the double mirror system is rotated around the intersecting edge the image does not move at all. If the intersecting edge is rotated or moved sidewise the image will move accordingly.


Y (Up)
X (North)
Projector (Observer alongside)
Mirror
P
West
Z (East)
South
Y
X
P
Z
θ
φ
ω
Up
Projected images reflected
North
West
East
y
z
Projected image not reflected
Down
Up
Y
Projected images reflected
East Z
West
Down

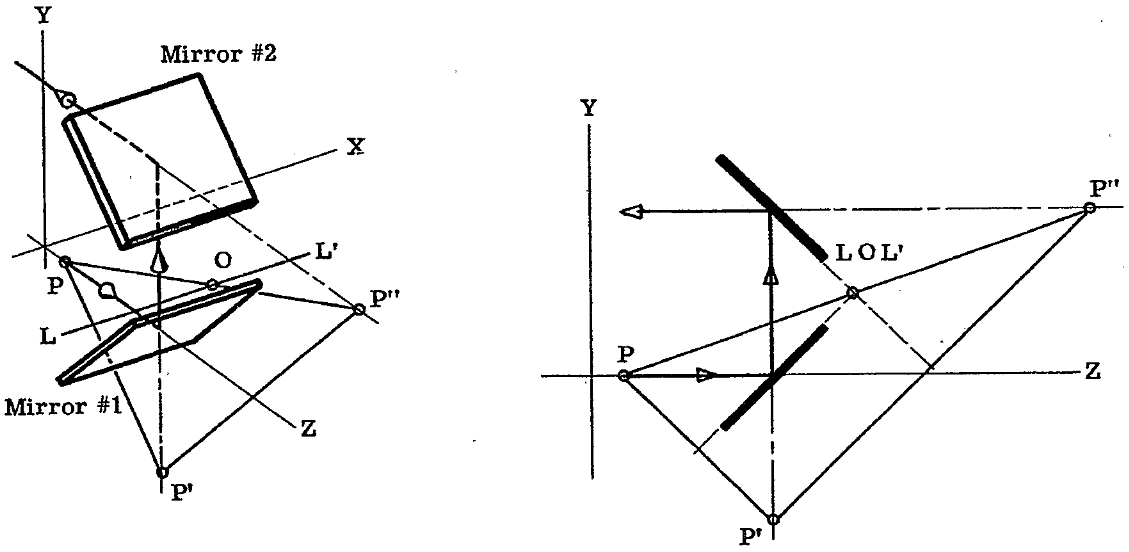
(a) The internal observer image sphere

(b) Geometry of image point.

(c) Outside sphere looking North

(d) Inside sphere looking South

Figure 13.15-The solid-angle image.



(a) Three dimensional view.

(b) Projection on YZ plane.

Figure 13.16-Reflection from two perpendicular mirrors.

13.7 TYPICAL PRISM SYSTEMS

13.7.1 Prisms and Mirrors. With the basic principles of mirror systems having been discussed, the analysis of some simple systems can be undertaken. In this analysis the reader should bear in mind that we are concerning ourselves principally with reflecting prisms. The reflecting faces of these prisms behave like mirrors rigidly mounted with respect to each other.

13.7.2 Illustration conventions.

13.7.2.1 In order to provide the reader with illustrations which require the minimum mental orientation to see both object and image correctly, we have portrayed the object as the letter R illuminated from behind by a collimated beam, the central ray of which is indicated by ——▷ . The image is illustrated by the appearance of the projected image that would be produced if a direct vision screen, such as frosted glass, were held normal to the emergent beam.

13.7.2.2 To observe either object or image the reader should view them as if the central ray from them were directed at his eye. When the limits of graphic art prohibit showing both object and image from the viewpoint of the observer, the projected image will be dashed to indicate it is shown from the wrong viewpoint. This enables illustration of the effect produced by multiple reflection systems without concern for the effect of each individual reflection. This does not permit indication of the apparent position of the virtual image (except for Figure 13.17 where both are shown) but does show left-or right-handedness.

13.7.3 The 45°-90°-45° Prism.

13.7.3.1 This simple prism can be used in many different ways. It can turn a beam through a 90 degree or 180 degree bend, or it can be used to invert an image.

13.7.3.2 To turn a beam through 90 degree, the prism is used as shown in Figure 13.17. Since there is only one reflection, the image is left-handed. The projected image is what the observer would see on a translucent back-lighted screen as described in paragraph 13.7.2, above. If the screen were removed, the virtual image would still be left-handed but located on the extended line of sight behind the reflecting surface as in the case of a single plane mirror. If the normal to the hypotenuse is in a horizontal plane the right hand object is swung around a vertical axis. The letter R will appear as Я . If the normal to the hypotenuse lies in the vertical plane the image will appear rotated around a horizontal axis. The letter R will appear as ꓤ .

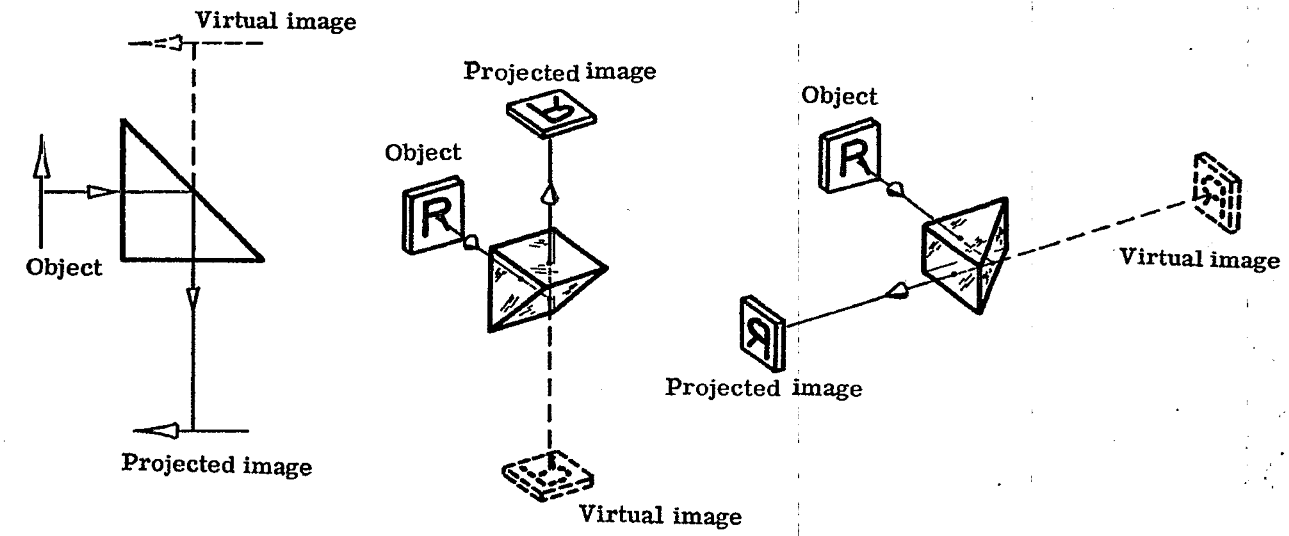


Figure 13.17-The 45° - 90° - 45° prism used as a right-angle prism.

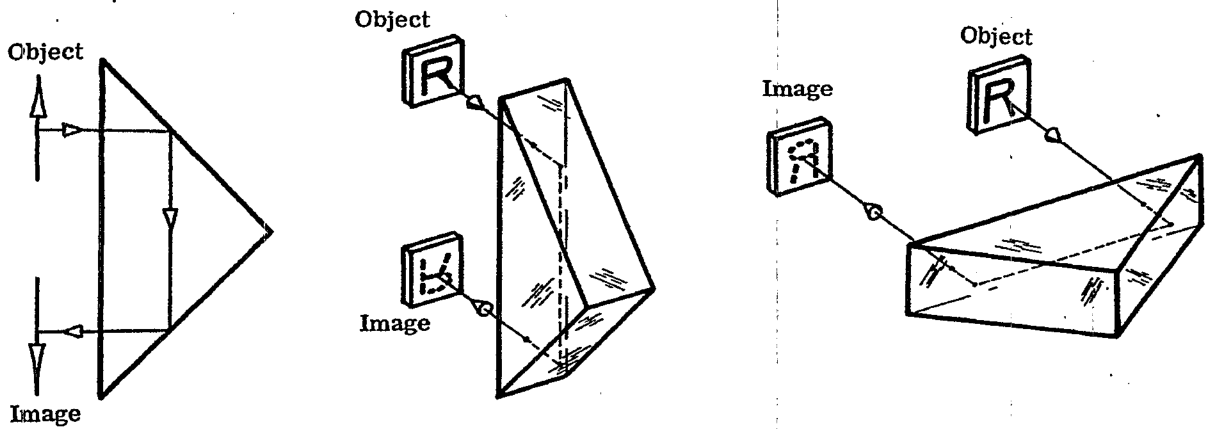


Figure 13.18-The 45° - 90° - 45° prism used as a Porro prism.

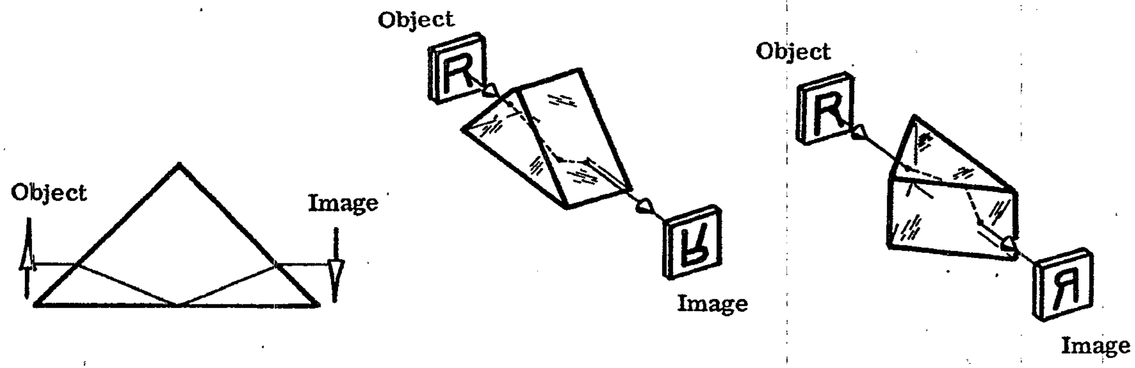


Figure 13.19-The 45° - 90° - 45° prism used as a Dove prism.

13.7.3.3 When used as a double mirror system, the prism is positioned as shown in Figure 13.18. Since there are two reflections, the image will be right-handed. In the illustration the projected image is shown in dashed lines indicating the observer would view it from the opposite side of the screen. To the observer so stationed it would appear as ꓤ if the roof edge of the prism (the edge formed by the intersection of the reflecting surfaces) is horizontal. If the roof edge is vertical, the image will appear as R. With this prism it is possible to rotate the image into any desired orientation and always have it right-handed. Used in this fashion, the 45° - 90° - 45° prism is called a Porro prism and will be discussed in detail later.

13.7.3.4 When used as shown in Figure 13.19, it is called a Dove prism and can be used to rotate an image. There is a single reflection so the image is left-handed. If the normal to the hypotenuse face lies in the vertical plane, the letter R appears as ʁ . When the normal lies in the horizontal plane, the image appears as Я

13.7.4 Use of prisms in telescope systems.

13.7.4.1 One of the main uses of prisms is to provide the proper orientation of the image in telescopes. The image in a simple telescope, which consists of an objective and eyepiece, is right-handed but upside down. The image may be made erect by using two prisms. In order to keep the image right-handed the prism system must have an even number of reflections. The minimum number of reflecting surfaces is two. A prism system which does this is shown in Figure 13.20 as it may be used in a telescope.

13.7.4.2 The prism illustrated in Figure 13.20 is called an Amici prism and is described in more detail in Section 13.7.5. It is essentially a 45° - 45° - 90° prism with the hypotenuse face made into a roof. It is for that reason often called a roof prism. In Figure 13.21 a beam is drawn showing how it reflects a cylinder of light. This drawing shows plan and elevation views of the prism. A view looking along the roof edge and a pictorial three dimensional view are also shown. The selected rays traced through the prism show how the image is rotated 180 degrees. The dotted lines show that this prism is cut out of a large Amici prism. One can see that as the cylinder of light passes through the prism the complete cylinder strikes first one face of the roof and then crosses over to the other roof. If the roof angle is not exactly 90 degrees the only effect is that the exit and entrance angles no longer remain in parallel planes. While permitting easier manufacturing tolerances, this method is seldom used because it requires too large a block of glass for the space and

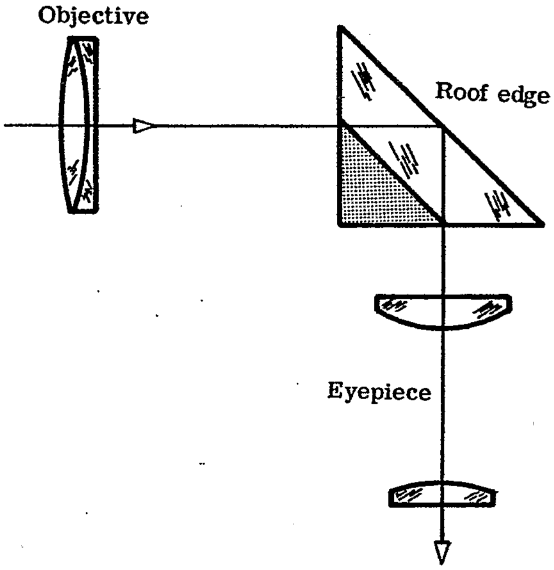


Figure 13.20-The Amici prism in a telescope.

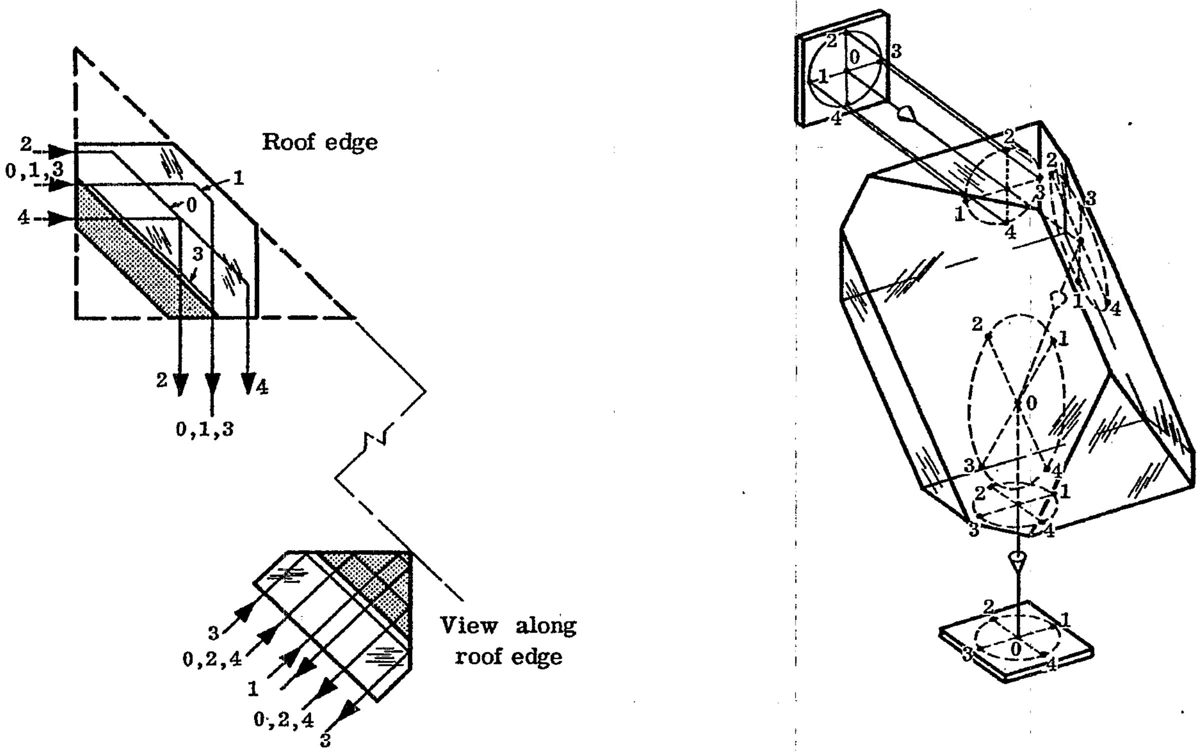


Figure 13.21-The Amici prism as a double reflector.

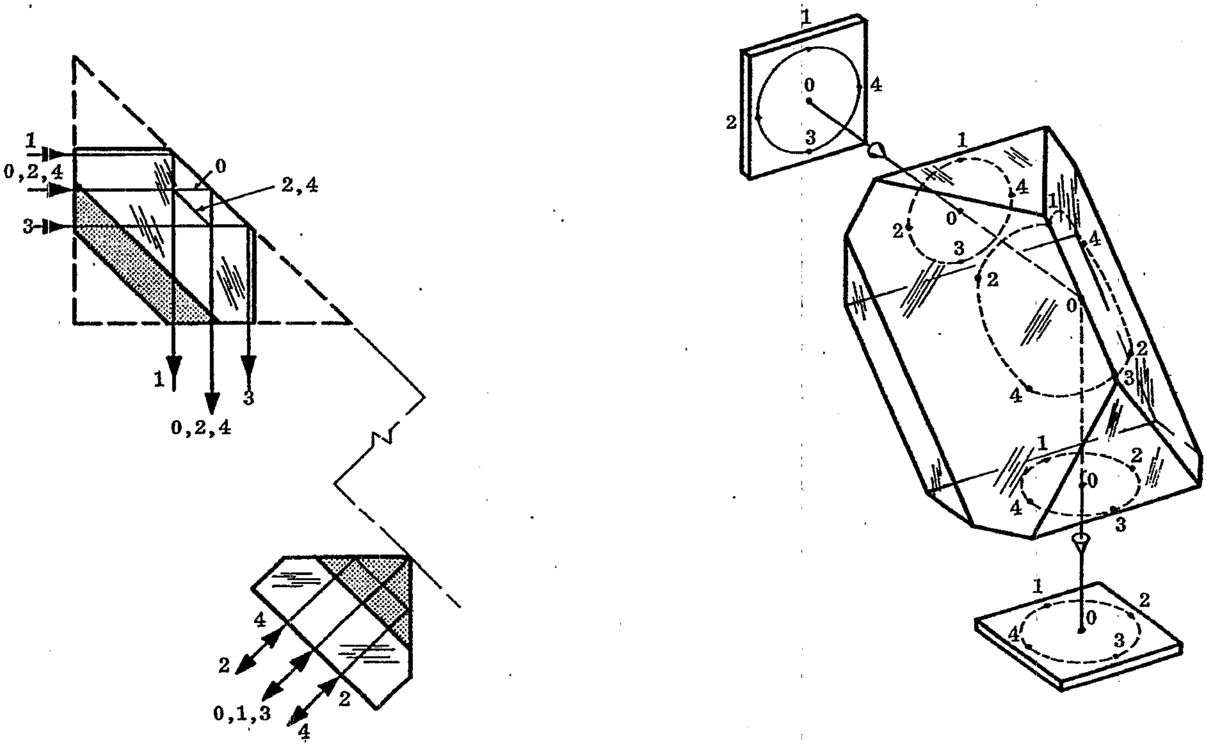


Figure 13.22-The Amici prism as a split reflector.

weight limitations of most applications.

13.7.4.3 A more common method of using the Amici prism is shown in Figure 13.22. This useage permits a much larger cylinder of light to pass through the same size prism, or conversely, to handle the same size cylinder of light with a much smaller prism than that of Figure 13.21. There is a fundamental difference between the two applications. In Figure 13.22 the beam is split by the prism's roof edge. If there is any error in the 90 degree roof edge angle the entering beam is split into two beams and a double image is formed. This means that if the Amici prism is used in this manner the 90 degree angle must be made to high degree of precision. In most applications this angle has to be held to 90 degree ± three or four seconds. Roof prisms as used in Figure 13.22 are very efficient as far as size goes but they are expensive to make because of the precision required. If the prism is used as shown in Figure 13.21 the accuracy required is not as high but the prism has to be much larger in order to pass the same size beam.

13.7.4.4 It is instructive to draw these views of roof prisms and show the path of rays passing through them. In Figure 13.21 the prism can be cut even further to reduce the weight of glass. How would one decide how it could be cut and not interfere with the beam passing through? In telescopes the objective is usually larger than the eyepiece field stop, so that the prism must pass a section of a cone rather than a cylinder. This means the entering circle and the exit circle are different sizes. It is a good exercise to try and lay out a prism of minimum size and then to determine how corners can be cut to further reduce the weight. This is prism design.

13.7.5 Prism rotation of the image through 180 degree. The Amici prism is the simplest method for erecting the image in a telescope, but it has the difficulty that one must look around a corner. A Dove prism with a roof on the hypotenuse face as shown in Figure 13.23 uses the double mirror principle of the Amici prism. This prism must be located in front of the objective in parallel light. If it is located in between the objective and the eyepiece it causes aberrations because of the refraction of the slanting surfaces. More will be said about this in the tunnel diagram Section 13.8. If it is necessary to have the optical axis of the telescope objective and eyepiece parallel, the Amici prism can be used with other prisms to bend the light through 90 degrees. It is necessary however to use two reflections in order to preserve the right-handed use of the image. Figure 13.24 shows a penta prism with two reflections which could be used with the Amici prism.

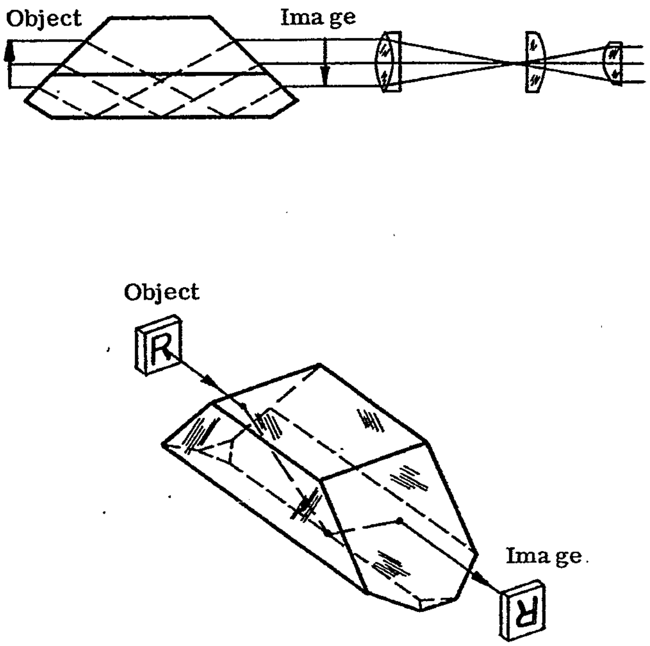


Figure 13.23-The Amici prism in telescope systems.

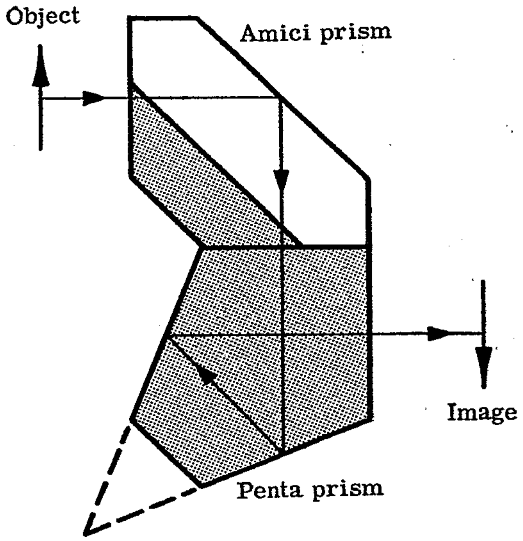


Figure 13.24-An Amici and penta prism system.

13.7.6 The Porro prism.

13.7.6.1 The most common method for erecting the image in a telescope is the Porro prism system. This is made up of two 45° - 45° - 90° prisms as indicated in Figure 13.25. The first prism is positioned so that the roof edge is perpendicular to the corresponding edge in the other prism. One can understand the action of the prism by considering the explanation illustrated in Figure 13.26.

13.7.6.2 This diagram illustrates one of the sources of confusion in understanding of prisms. It shows how the prism A rotates the image around the line of the intersecting edge. But note that as shown the R is left-handed. Why is this when it has been clearly stated that double reflection always provide a right-handed image? If the reader will recall Paragraph 13.7.2 on illustration conventions, it then will become apparent that in the drawing of Figure 13.26 the image of the object is not being presented from the viewpoint of the observer. If you imagine standing and looking at the original object, then it would not be possible to see the image after passing through prism A. It would be necessary to turn yourself completely around. The image shown in the drawing is the image as viewed with the light moving away from you. If you turn yourself around and look at this image from the back of the paper it will appear right-handed. Prism B in effect does this for us. It merely reflects the image from prism A around so that it can be seen from the same direction as the original object.

13.7.6.3 Figure 13.26 shows that the orientation of the final image depends only on the relative positions of the intersecting edges of the two prisms. As long as they are perpendicular to each other the final image is completely erected. If there is an error from perpendicularity of the amount ϵ, the image will be rotated by 2ϵ.

13.7.6.4 The Porro system is a popular design because the 45° - 45° - 90° prisms can be made with reasonably broad tolerances in the angles. The optical beam is not split as it is with the roof prism so prism angle errors do not cause any image doubling. Angle errors merely cause a deviation in the optical axis as it passes through the prism. The exit optical axis may not end exactly parallel to the entering axis.

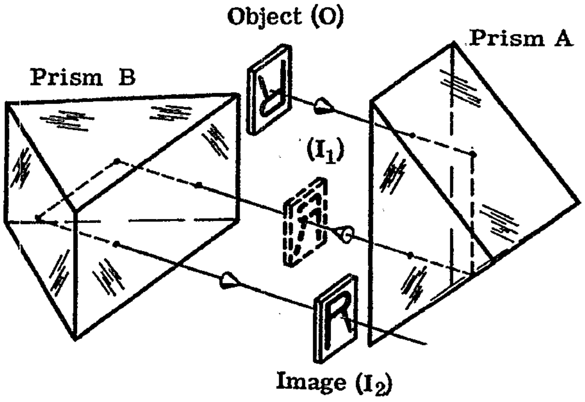


Figure 13.25-Reflections through the Porro prism.

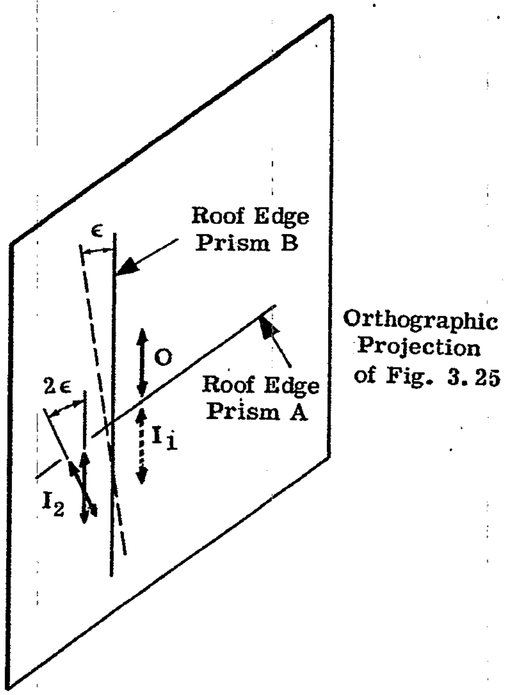


Figure 13.26-Image rotation in the Porro prism.

13.8 THE TUNNEL DIAGRAM

13.8.1 Right angle prism tunnel.

13.8.1.1 It is very convenient in laying out prisms to "fold" the prism around the reflecting surfaces. This generates a tunnel diagram. Consider the prism in Figure 13.27 The hypotenuse face BC can be considered as a mirror. The faces AB and AC may be considered as imaged in this mirror as shown dotted. The ray of light passing through the prism may also be considered imaged as shown. An observer looking into face AB therefore sees face AC at A'C. It appears as though he is looking straight through a block of glass of thickness BA'. One can check immediately that the angle ABC is equal to the angle ACB then the imaged face A'C is parallel to the face BA. Optically then the prism introduces a block of glass in the optical system. As far as design considerations are concerned the prism may be considered as merely the insertion of a thick block of glass and may be treated as two ordinary parallel plane surfaces where rays are traced as straight lines within the prism.

13.8.1.2 The tunnel diagram helps one to realize that any prism system used to erect images or turn light around corners should "fold" out in a tunnel diagram so that the entering and exit faces are parallel. If they end up nonparallel then the prism will cause chromatic dispersion.

13.8.2 The Porro prism tunnel.

13.8.2.1 Figure 13.28 is the tunnel diagram for the 45° - 45° -90° prism as used in a Porro system. The original Porro, ABC, Figure 13.29, has been folded around AB to image C as C' and around BC' to image A as A'. The tunnel diagram, Figure 13.28, is then a square with AC', A'C' and A'C as images of AC while A'B and BC' are images of AB and BC respectively. However, since the prism is now considered to be replaced by a glass block and since AB, BC, A'B and BC' all lie within the block, we can ignore them, as a little thought will soon show. We can now easily lay out rays entering the block through face AC by computing their refraction and extending the refracted ray on a straight line through the prism.

13.8.2.2 Let us consider the passage of several rays traced through the Porro prism Figure 13.29 and through the tunnel diagram Figure 13.28. Ray R_1 Figure 13.29 enters parallel to and above the optical axis of the prism and is reflected parallel to and an equal distance below the optical axis as R'_1. In the tunnel diagram,

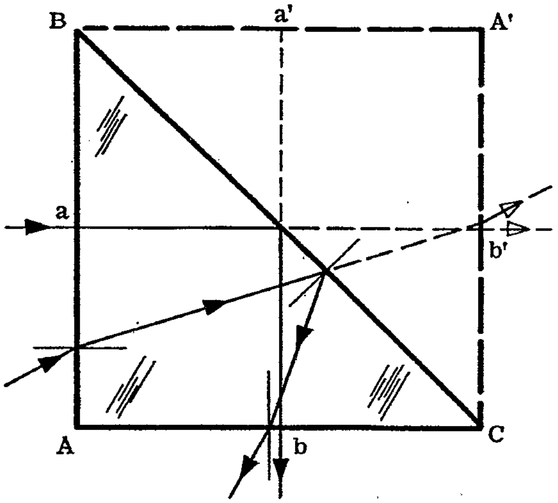


Figure 13.27-Right-angle prism tunnel diagram.

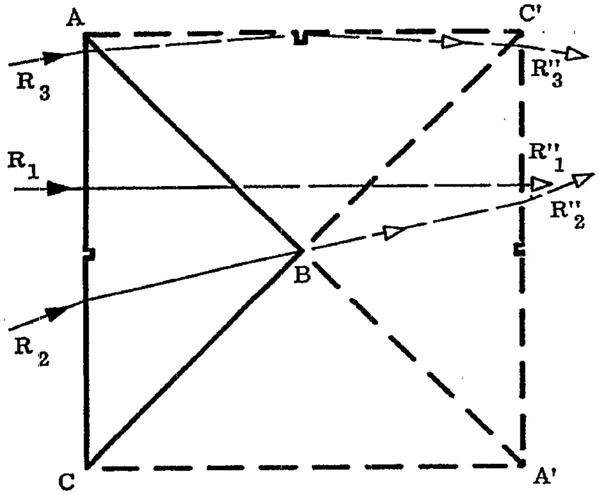


Figure 13.28-Porro prism tunnel diagram.

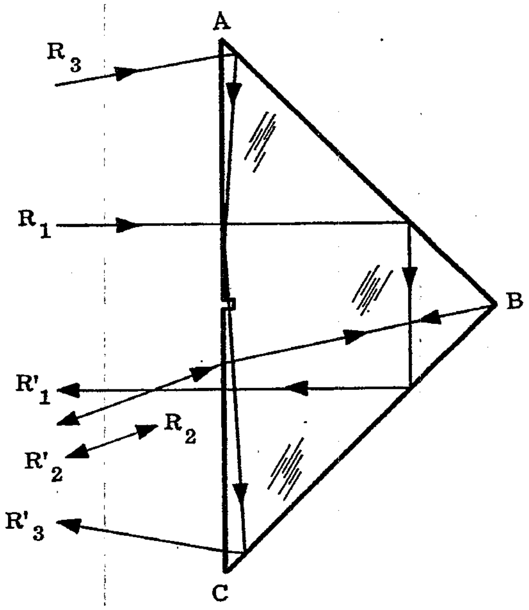


Figure 13.29-The Porro prism.

it emerges as R''_1 , above the optical axis. However, note its relation to A' and C' as compared with the R'_1 relation to A and C. This tells us that the designer must interpret the tunnel diagram in the light of his knowledge of prism effect on the image orientation.

13.8.2.3 Consider further the ray R_2 in Figure 13.29 which enters the prism so as to strike the roof edge and be reflected back upon itself. Note the path of R''_2 in Figure 13.28 and again observe relation to A'C'.

13.8.2.4 The tunnel diagram is particularly useful in detecting the presence of unwanted reflections. In Figure 13.28 notice the ray R_3 entering the prism near A. It passes through the tunnel diagram very close to the hypotenuse face. A slight inclination of this ray and it could reflect off the hypotenuse surface as shown by the ray R''_3 . This ray encounters three reflections in passing through the prism. This would cause a left-handed image. Since the prism is intended to be used with two reflections these rays with the extra reflection are called ghost rays. The ghost reflections may be eliminated by cutting a notch in the prism as shown in Figure 13.28. The tunnel diagrams for several prisms are shown in the data sheets on prisms at the end of this section.

13.8.3 The reduced or apparent prism length.

13.8.3.1 We have now satisfied ourselves that when prisms are introduced into an optical system they behave optically as would a block of solid glass with plane parallel faces; that rays may be easily traced through by refracting at entrance and exit faces, with the refracted ray travelling in a straight line within the prism; that the entering and exiting ray will be parallel. Consider then the point P on the surface of the block of glass shown in Figure 13. 30. By using equations 6 - (2), 6 - (3), and 6 - (4), it may easily be determined that the image P' lies at a distance $t\,\frac{(n-1)}{n}$ from P. This means that from the right hand side of the block, P appears to be separated by t/n surface of the prism, or the prism appears to have a thickness of t/n. This is variously called the reduced or apparent thickness of the prism or the air-equivalent prism.

13.8.3.2 In drawing tunnel diagrams it is convenient to draw the reduced tunnel diagram. The actual and the reduced tunnel diagram for a penta prism are shown in Figure 13.31. The reduced prism is convenient for it is possible to trace rays directly through it without refracting them at the outside surfaces. This is of course an approximation since the effective thickness of a block was computed to be t/n with paraxial ray approximations.

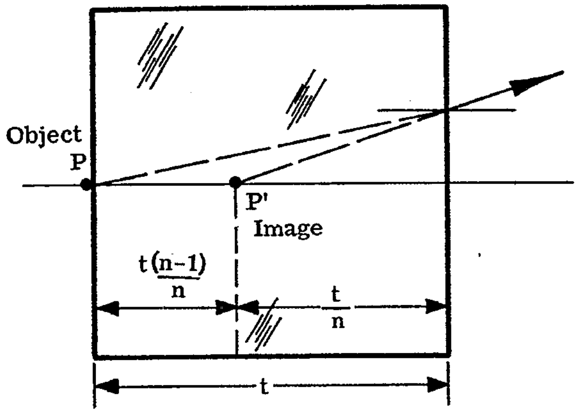


Figure 13.30-The apparent thickness of a glass block.

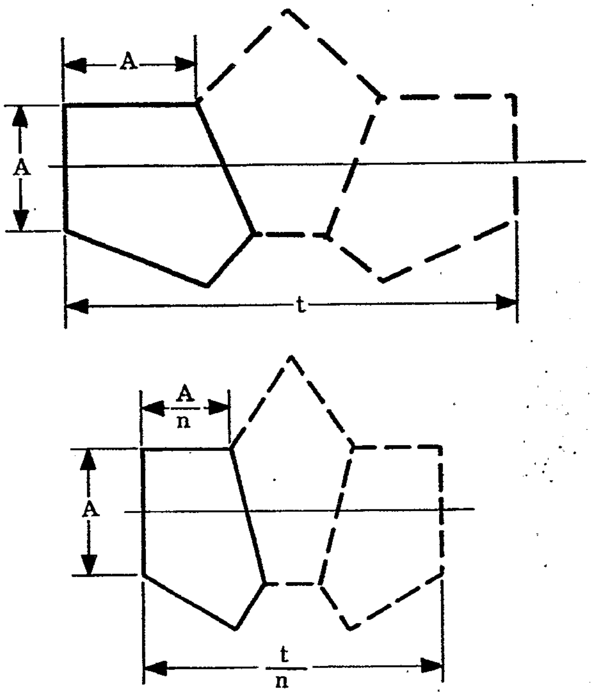


Figure 13.31-Actual and reduced tunnel diagram of a penta prism.

13.9 ABERRATIONS INTRODUCED BY PRISMS

13.9.1 Typical orientation. Reflecting prisms are generally designed so that the entering and exit faces are parallel and the entrance face is perpendicular to the optical axis. The aberrations introduced by the block of glass so oriented may be corrected by the normal centered lens system. The prism adds aberrations however only if it is located in a convergent or divergent beam of light. If the prism is in parallel light which is perpendicularly incident on the entrance or exit face, obviously no refraction, and therefore no aberrations will be introduced.

13.9.2 The third order aberrations introduced by a prism of thickness t and index n.

13.9.2.1 Figure 13.32 shows a block of glass in a convergent beam of light. The third order calculations for B, F and C are included in Table 13.1. The contributions to E, a and b, are not included. They may be readily calculated as an exercise. One should notice that the total aberrations introduced do not depend on y_1 or $\bar{y}_1$. This shows that, as long as its faces are perpendicular to the optical axis of the system, the position of the prism has no influence on the aberrations. If the optical axis of the prism is parallel to but displaced from the system's axis, occlusion of part of the beam may occur with the resultant loss of imagery being comparable to the effect of an unsymmetrical stop being introduced. Angular misalignment however, will have the effect of changing the value of t and, further, will introduce assymmetry into the system.

13.9.2.2 The problem of prism design then, is not complete until the designer has computed manufacturing tolerances on the prism faces and provided for proper alignment within the system. Fortunately the latter is usually a problem in line with centering the instrument system, while the former is somewhat simplified by existence of design data on many commonly used prisms. This data is presented in the remaining pages of this section.

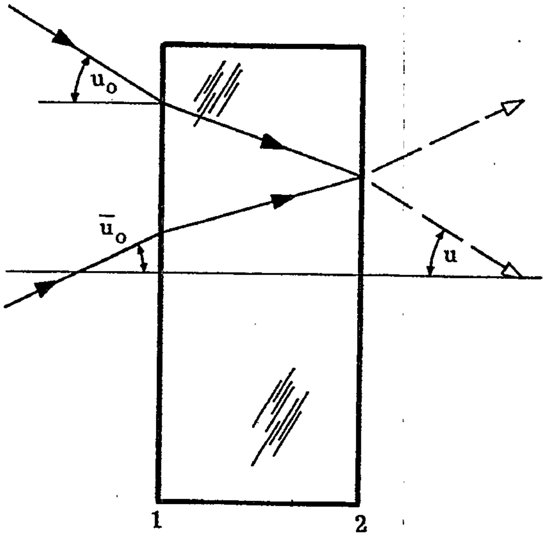


Figure 13.32-The glass block and the convergent beam of light.

Surface	0		1		2		Totals
c			0		0		
t		1		t		2	
n		1		n		1	
$\left(\frac{n_{-1}}{n} - 1\right)$			$\left(\frac{1}{n} - 1\right)$		$(n-1)$		
y			y_1		$(y_1 - \frac{t}{n} u_o)$		
u		u_o		$u_{o/n}$		u_o	
i			u_o		$u_{o/n}$		
$\bar{y}$			$\bar{y}_1$		$\bar{y}_2$		
$\bar{u}$		$\bar{u}_o$		$\bar{u}_{o/n}$		$\bar{u}_o$	
$\bar{i}$			$+\bar{u}_o$		$\bar{u}_{o/n}$		
S			$-y_1 u_o (n^2 - 1)/n^2$		$y_1 u_o (n - 1) - \frac{t}{n} u_o^2 (n^2 - 1)$		
B			$-y_1 u_o^3 \frac{(n^2-1)}{n^2}$		$y_1 u_o^3 \frac{(n^2-1)}{n^2} - \frac{t}{n^3} u \frac{(n^2-1)}{n^2}$		$\Sigma B = -t u_o^4 \frac{(n^2-1)}{n^3}$
F			$-y_1 u_o^2 \bar{u}_o \frac{(n - 1)}{n}$		$y_1 u_o^2 \bar{u}_o \frac{(n^2-1)}{n^2} - t u_o^3 \bar{u}_o \frac{(n^2-1)}{n^3}$		$\Sigma F = -t u_o^3 \bar{u}_o \frac{(n^2-1)}{n^3}$
C			$-y_1 u_o \bar{u}_o^2 \frac{(n^2-1)}{n^2}$		$y_1 u_o \bar{u}_o^2 \frac{(n^2-1)}{n^2} - t u_o^2 \bar{u}_o^2 \frac{(n^2-1)}{n^3}$		$\Sigma C = -t u_o^2 \bar{u}_o^2 \frac{(n^2-1)}{n^3}$

Table 13.1-The third order aberrations introduced by a prism.

13.10 PRISM DATA SHEETS

13.10.1 Introduction.

13.10.1.1 The prism data sheets are presented as a guide to the designer and provide him with an orthographic projection, a tabular list of the dimensions, a tunnel diagram and a brief description of many different kinds of prisms.

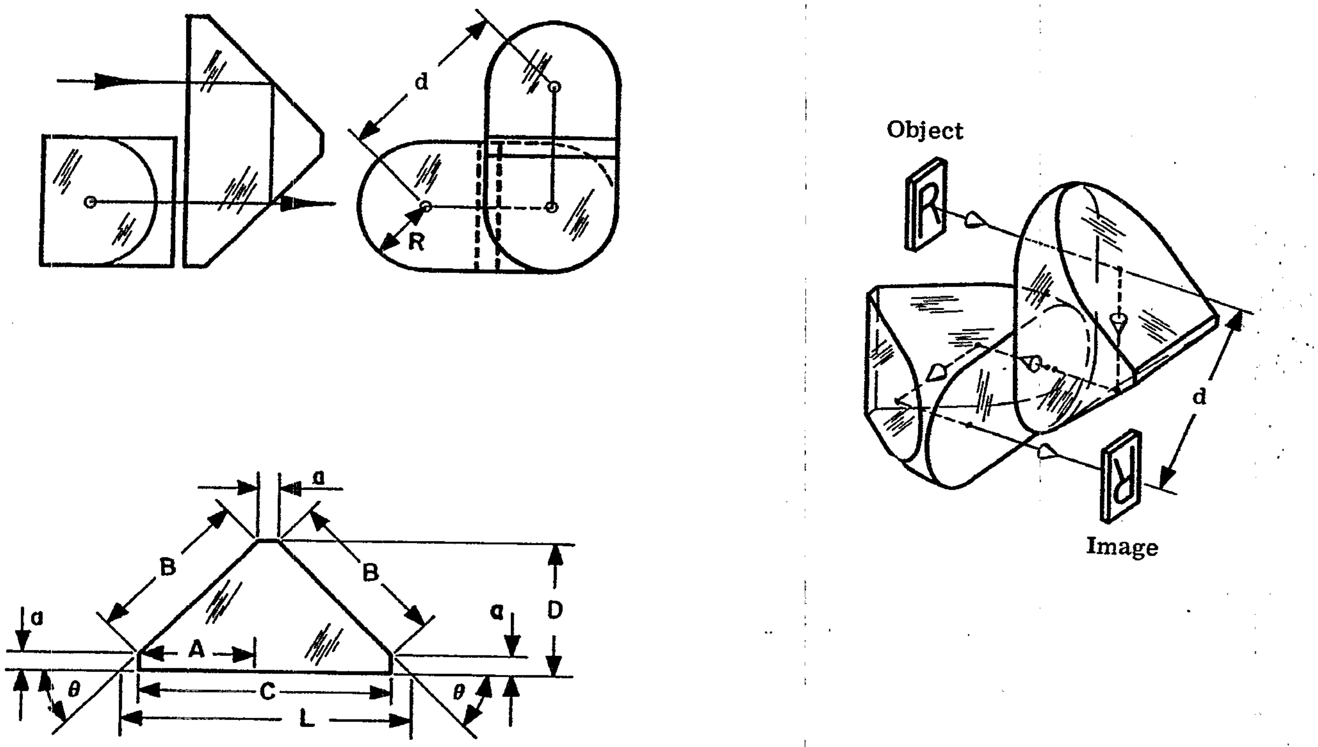
13.10.1.2 Notice that in the following data sheets, the terms invert and revert are used to describe the image. Invert means to rotate the object plane about a horizontal line in or parallel to the plane and produces the left-handed image one sees in a reflecting pool. Thus, for object R, the inverted image is ʁ. Revert means to rotate the object plane about a vertical line in or parallel to the plane and produces the left-handed image one sees in a shaving or dressing mirror. Thus, for the object R, Я is a reverted image. Obviously then for the object R, ᴚ is an inverted and reverted image.

13.10.1.3 The term "displace" refers to parallel separation of two lines. Thus we find that if an oblique ray strikes the entrance face of a plane parallel block, the ray leaving the exit face is parallel to but displaced from the entering ray. The word "deviate" refers to an angular relation between two lines. Thus in the foregoing example the line tracing the ray through the block is deviated by refraction at the surface.

13.10.1.4 The following symbols are used on the prism data sheets:

NOTATION	USE
Lettering guide capitals A, B, C,...	Linear dimensions of the geometric figure.
L	Over all length.
Lettering guide lower case a, b, c,...	Dimensions which are trigonometric functions of corresponding capital letters.
d	Displacement of the axial ray.
t	Optical path length of axial ray.
n	Index of refraction of the glass.
Greek letters	Angles.
α, β, γ	Direction angles.

13.10.2 Porro Prism System. In 1850 the Italian engineer Porro designed the prism system discussed here. This system consists of two right-angle prisms, usually identical in construction, placed at right angles to each other. It is a direct vision prism system but the axis is displaced by the amount d. This system will invert and revert the image.



$A = 1.00$ $n = 1.5170$ $\theta = 45°$ (These values are given) $a = 0.10$ (chosen arbitrarily)

$R = A/2 = 0.50$ $B = 1.4142A = 1.4142$ $C = 2A + a = 2.1$ $D = A + a = 1.1$

$L = 2A + 3a = 2.30$ $d = 1.4142\,(A + a) = 1.5556$ $t = 2\,(2A + 3a) = 4.60$ $t/n = 3.0324$

Figure 13.33-Porro prism system.

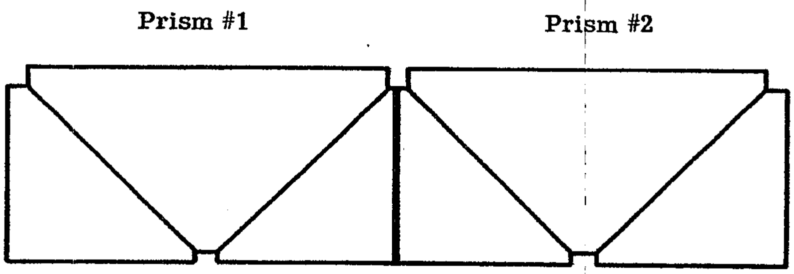
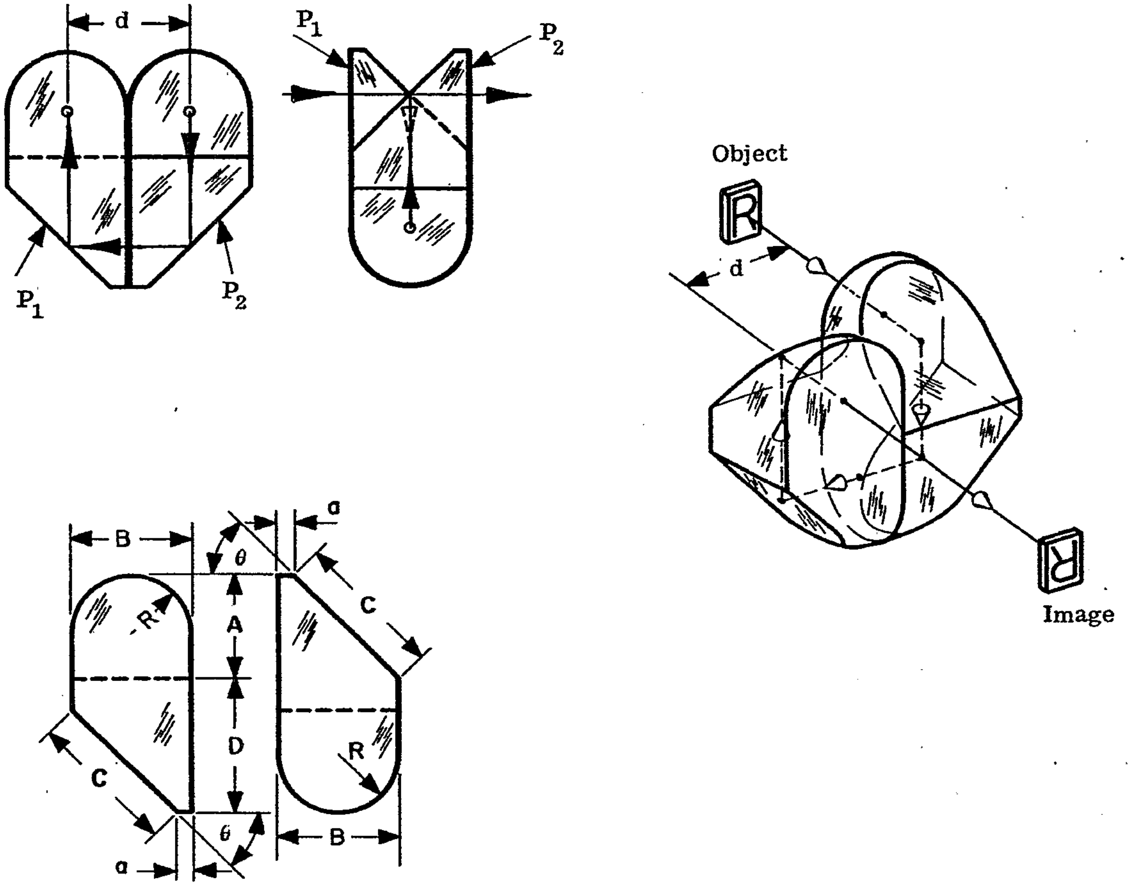


Figure 13.34-Porro prism tunnel diagram.

13.10.3 Abbe's Modification of the Porro Prism System. This prism system consists of two prisms cemented together. It will invert and revert the image. The system is a direct vision prism but the line of sight will be displaced by the amount d.



$A = 1.00$ $n = 1.5170$ $\theta = 45°$ $a = 0.10$ (chosen arbitrarily) $B = A + a = 1.10$

$C = 1.4142A = 1.4142$ $D = A + 2a = 1.20$ $R = B/2 = 0.55$ $d = B = 1.10$ $t/n = 3.0323$

$t = 2(2A + 3a) = 4.60$

Figure 13.35-Abbe prism system.

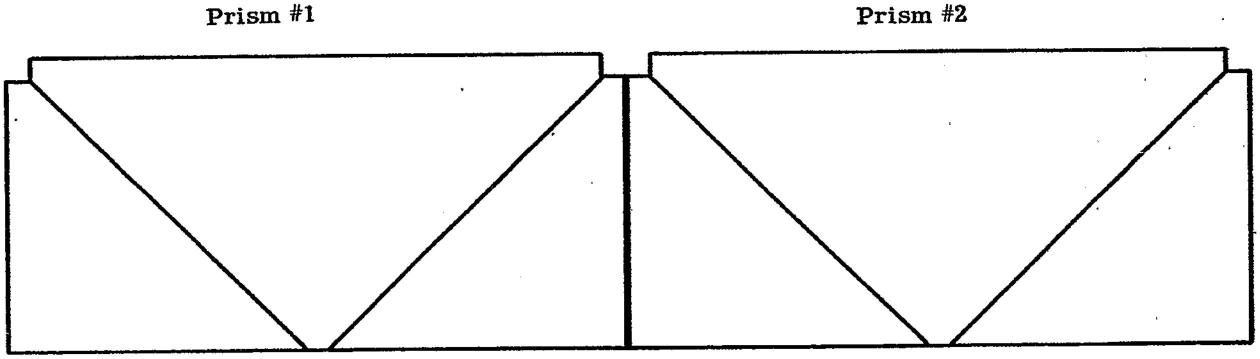
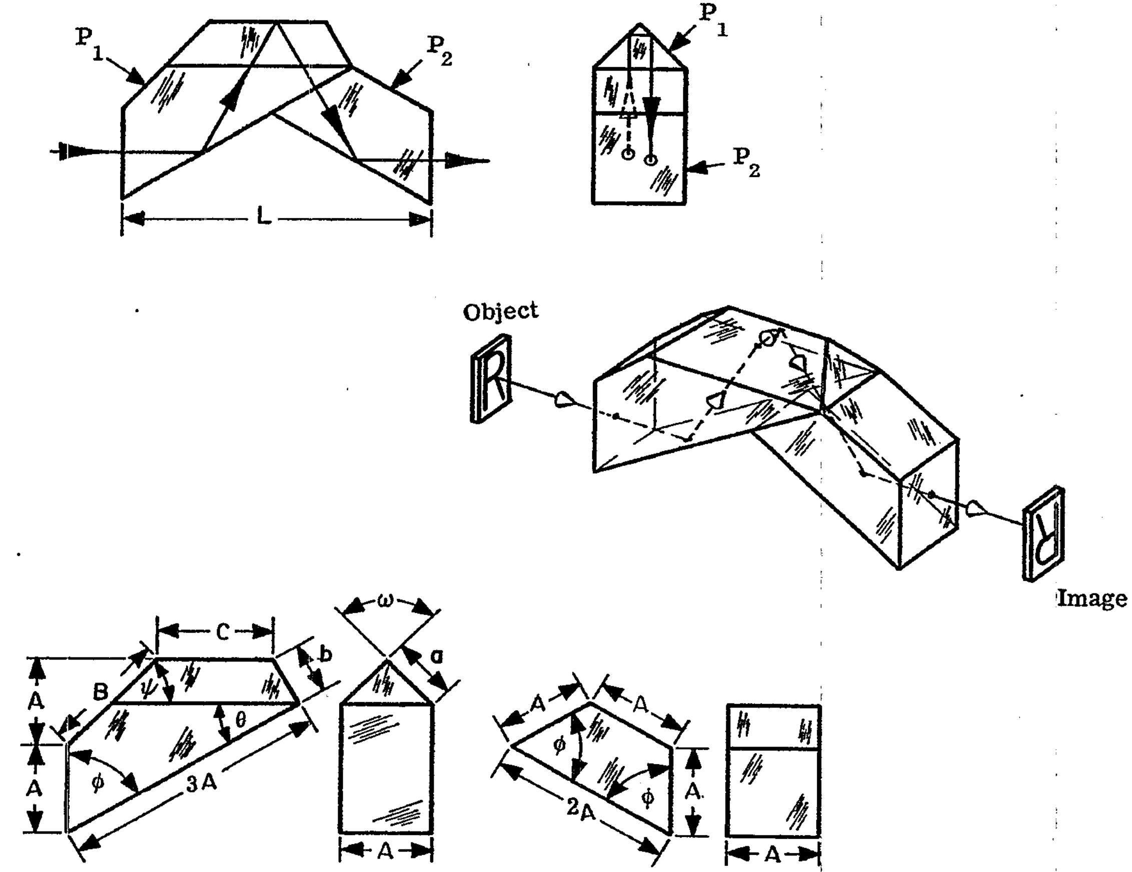


Figure 13.36-Abbe prism system tunnel diagram.

13.10.4 Abbe Prism, Type A. This prism inverts and reverts the image, but will not deviate the line of sight; hence, it is a "Direct Vision Prism." The prism is made in two pieces which are cemented together.



A = 1.00 θ = 30° ω = 90° n = 1.5170 ϕ = 60° ψ = 45° $B = 1.4142A$ = 1.4142
$C = 1.3094A$ = 1.3094 $a = 0.7071A$ = 0.7071 $b = 0.5774A$ = 0.5774 $L = 3.4644A$ = 3.4644
$t = 5.1962A$ = 5.1962 t/n = 3.4253

Figure 13.37-Abbe prism, type A.

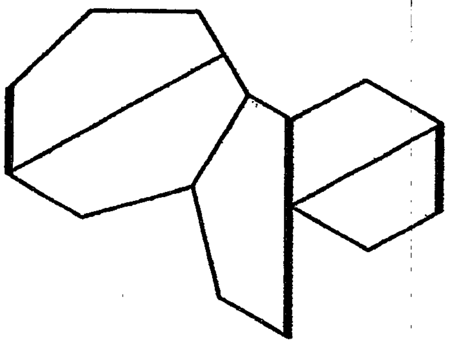
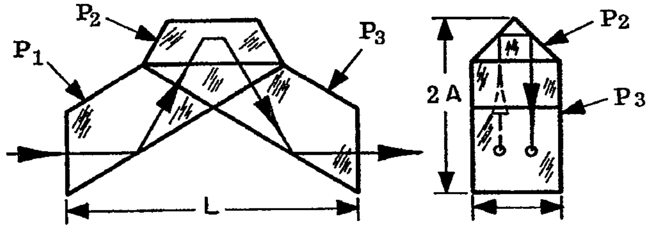
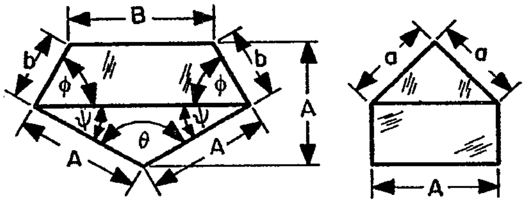


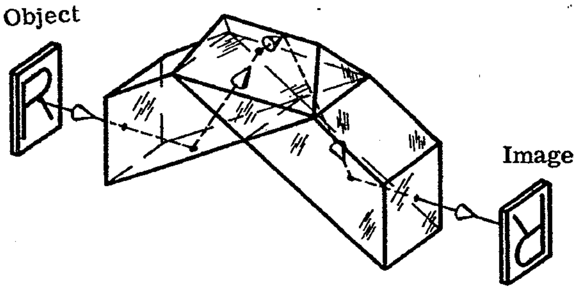
Figure 13.38-Abbe prism, type A, tunnel diagram.

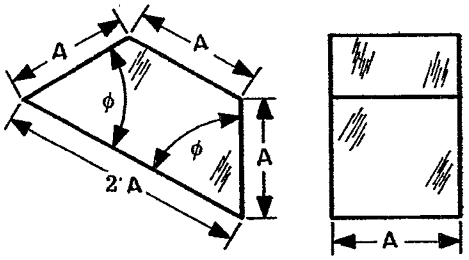
13.10.5 Abbe Prism, Type B. This prism is made of three single units which are cemented together. This prism will invert and revert the image but will not deviate the line of sight. This also is a "Direct Vision Prism."





Prism #2





Prism #1 and 3

$A = 1.00$ $\theta = 135^\circ$ $\omega = 45^\circ$ $\phi = 60^\circ$ $\psi = 30^\circ$ $n = 1.5170$ $a = 0.7071A = 0.7071$ $t/n = 3.4253$
$b = 0.5773A = 0.5773$ $B = 1.1547A = 1.1547$ $L = 3.4641A = 3.4641$ $t = 5.1962A = 5.1962$

Figure 13.39-Abbe prism, type B.

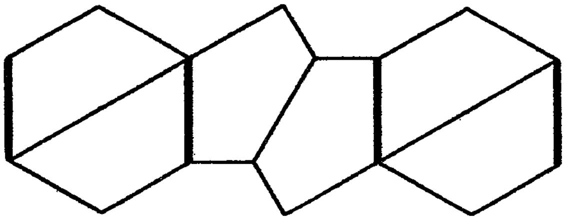
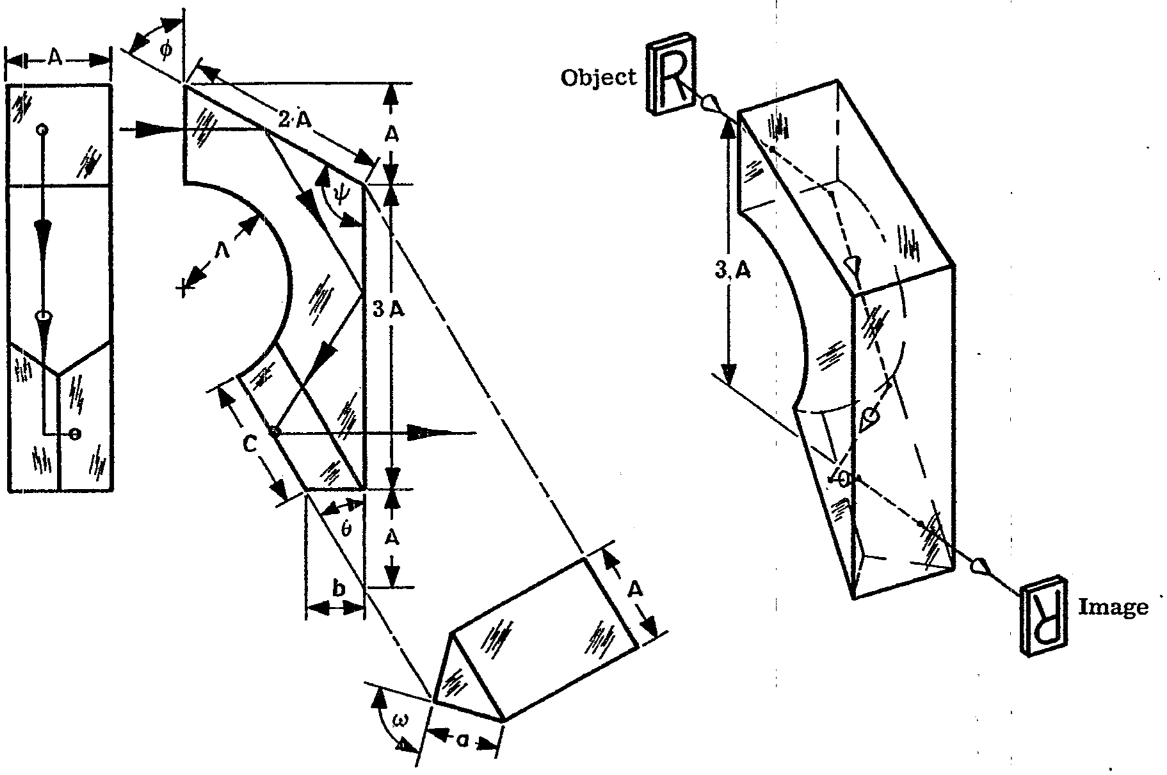


Figure 13.40-Abbe prism, type B, tunnel diagram.

13.10.6 Leman Prism. The Leman prism will revert and invert the image. The line of sight will be displaced laterally by an amount equal to 3A inches .



$A = 1.00$ $B = 1.7321A = 1.7321$ $n = 1.5170$ $a = 0.7071A = 0.7071$ $\theta = 30°$ $C = 1.3099A = 1.3099$
$\phi = 60°$ $b = 0.5774A = 0.5774$ $\omega = 90°$ $\psi = 120°$ $t = 5.1962A = 5.1962$ $t/n = 3.4253$

Figure 13.41-Leman prism.

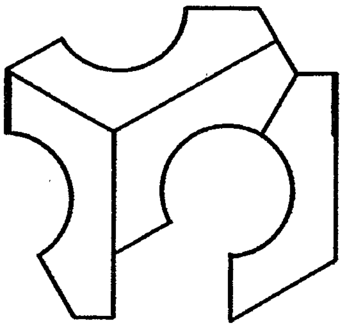
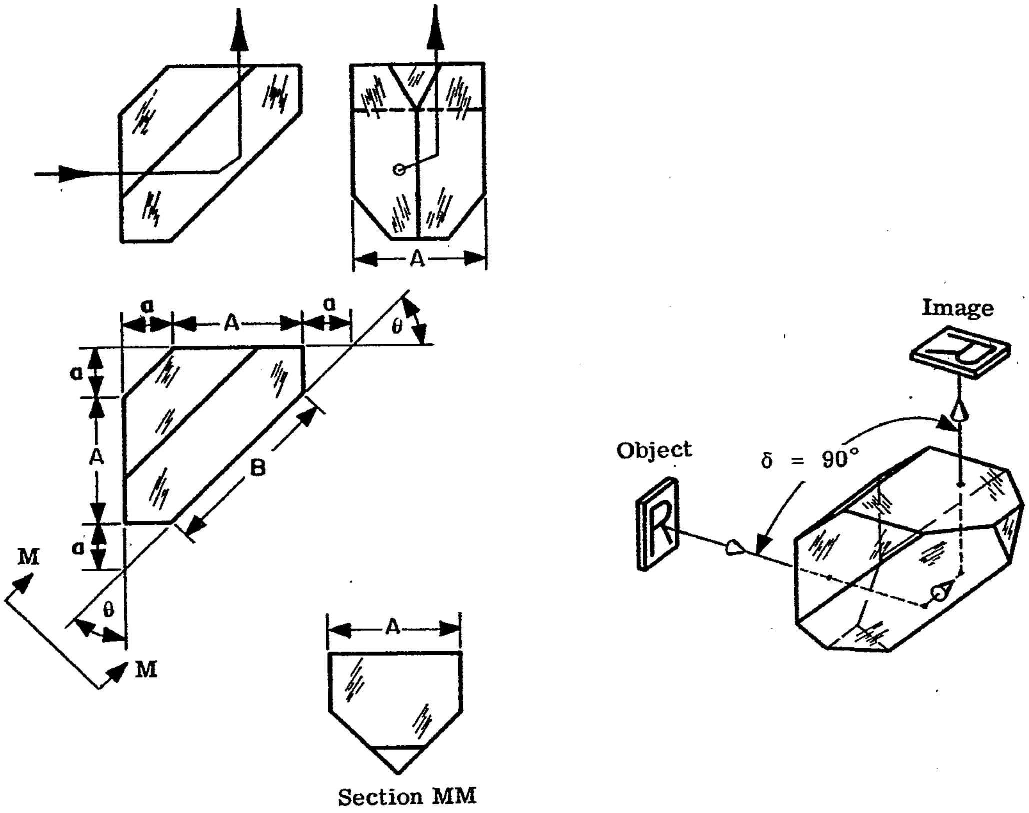


Figure 13.42-Leman prism tunnel diagram.

13.10.7 Amici Prism. During his life, 1784 to 1863, the Italian astronomer Amici designed many prisms. This is one of them. This prism will revert and invert the image and, at the same time, it will deviate the line of sight through an angle δ of 90°.



A = 1.00 n = 1.5170 θ = 45° **B** = 1.4142**A** = 1.4142 **a** = 0.3536**A** = 0.3536 t/n = 1.1253
t = 1.7071**A** = 1.7071

Figure 13.43-Amici prism.

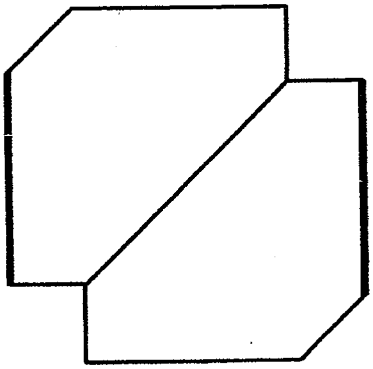
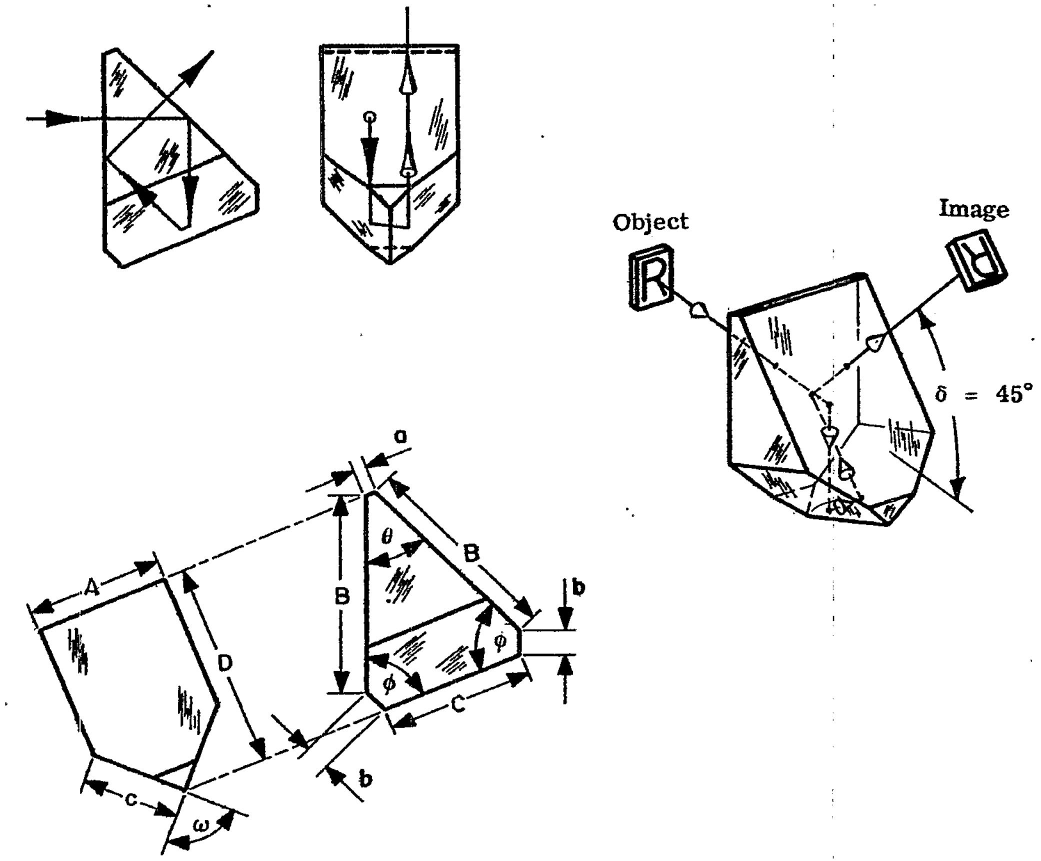


Figure 13.44-Amici prism tunnel diagram.

13.10.8 Schmidt Prism. This prism will revert and invert the image and, at the same time, it will deviate the line of sight through an angle $\delta = 45°$.



$A = 1.00$ $n = 1.5170$ $\theta = 45°$ $\omega = 90°$ $\phi = 67° 30'$ $a = 0.10$ (chosen at will) $t/n = 2.2506$

$B = 1.4142A + 0.5412a = 1.4683$ $C = 1.0824A = 1.082$ $D = 1.4142A + 2.3890a = 1.6531$

$t = 3.4142A = 3.4142$ $c = 0.7071A = 0.7071$ $b = 1.8478a = 0.1848$

Figure 13.45-Schmidt prism.

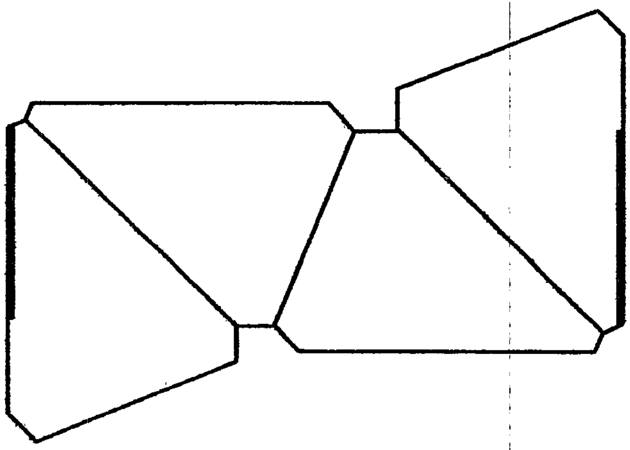
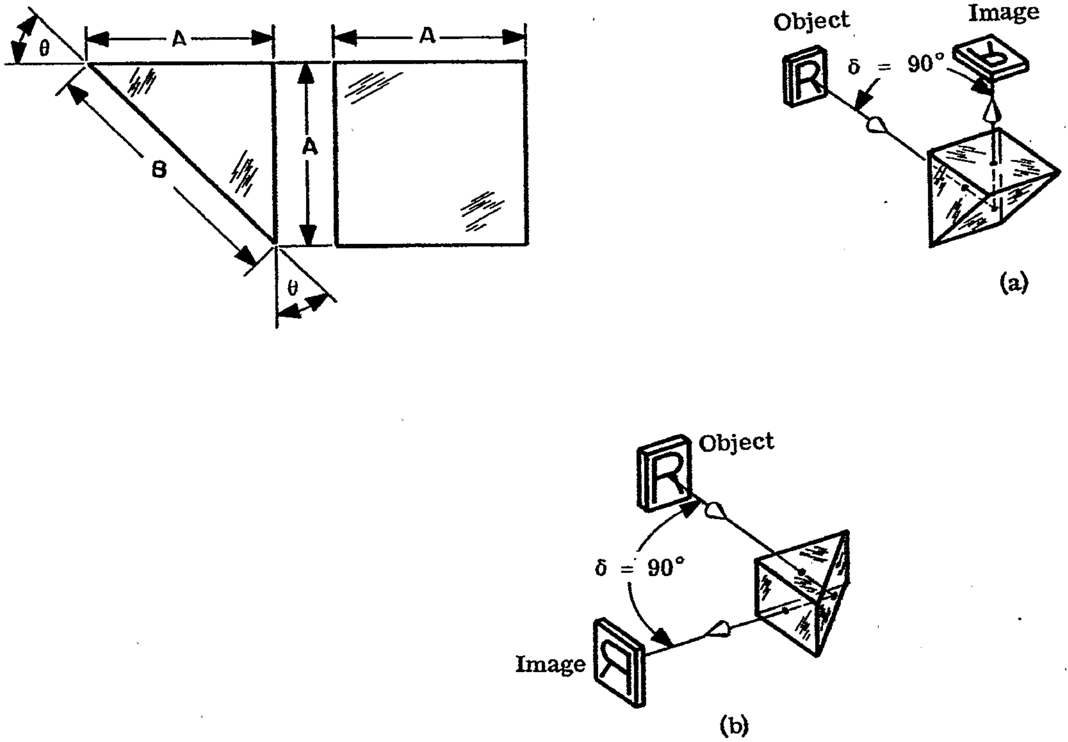


Figure 13.46-Schmidt prism tunnel diagram.

13.10.9 Right-Angle Prism. This single prism will deviate the line of sight through an angle $\delta = 90°$. The image will be inverted when the prism is held before the eye as shown in Figure 13.47(a), and it will appear reverted when the prism is turned through an angle of 90° as illustrated in Figure 13.47(b).



$A = 1.00$ $n = 1.5170$ $\theta = 45°$ $B = 1.4142A = 1.4142$ $t = A = 1.00$ $t/n = 0.6592$

Figure 13.47-Right-angle prism.

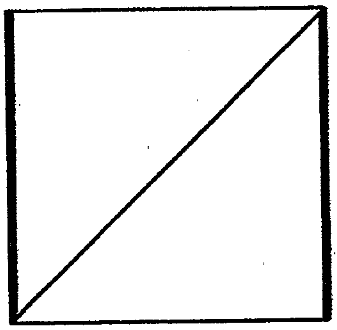
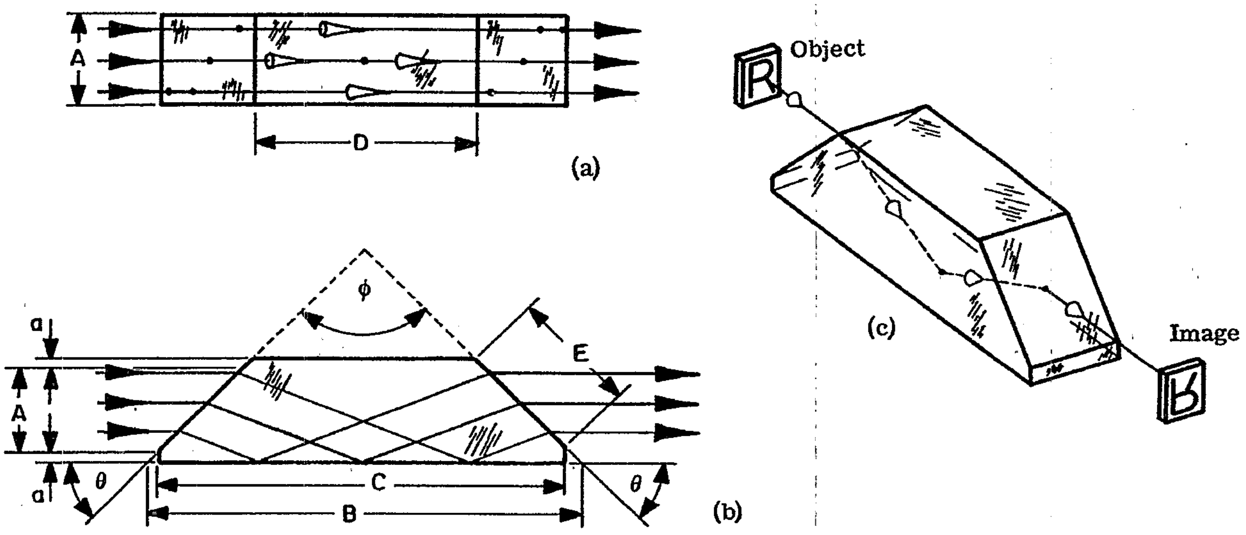


Figure 13.48-Right-angle prism tunnel diagram.

13.10.10 Harting-Dove Prism. This direct vision prism is made in one piece. The image will be inverted when the prism is held as shown in Figure 13.49(c), and it is inverted when the prism is turned about the axis through an angle of 90°. It can be used only in parallel light.



Effect on the Prism Constants When Different Types of Glass are Used

n =	1.5170	1.5725	1.6170	1.7200
B =	4.6498	4.4303	4.2822	4.0072
C =	4.5498	4.3303	4.1822	3.9072
D =	2.4498	2.2303	2.1822	1.9072
E =	1.4849	1.4849	1.4849	1.4849
t =	3.7165	3.5071	3.3637	3.1084
t/n =	2.4499	2.2303	2.0802	1.8072

$A = 1.00 \quad a = 0.05 \quad \phi = 90° \quad \theta = 45° \quad D = B - 2\,(A + 2a) = 2.4498 \quad n = 1.5170 \quad t/n = 2.4499$

$$B = (A + 2a)\left[\frac{\sqrt{n^2 - \sin^2\theta} + \sin\theta}{\sqrt{n^2 - \sin^2\theta} - \sin\theta} + 1\right] = 4.2271\,(A + 2a) = 4.6498 \qquad C = B - 2a = 4.5498$$

$$t = \frac{n\,(A + 2a)}{\sin\theta\,\sqrt{n^2 - \sin^2\theta} - \sin\theta} = 3.3787\,(A + 2a) = 3.7165 \qquad E = \frac{a + A}{\cos\theta} = 1.4142\,(A + 2a) = 1.4849$$

Figure 13.49-Dove prism.

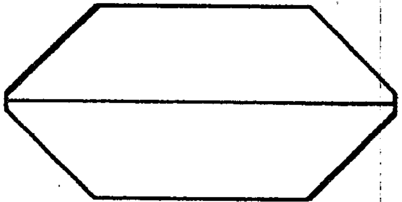
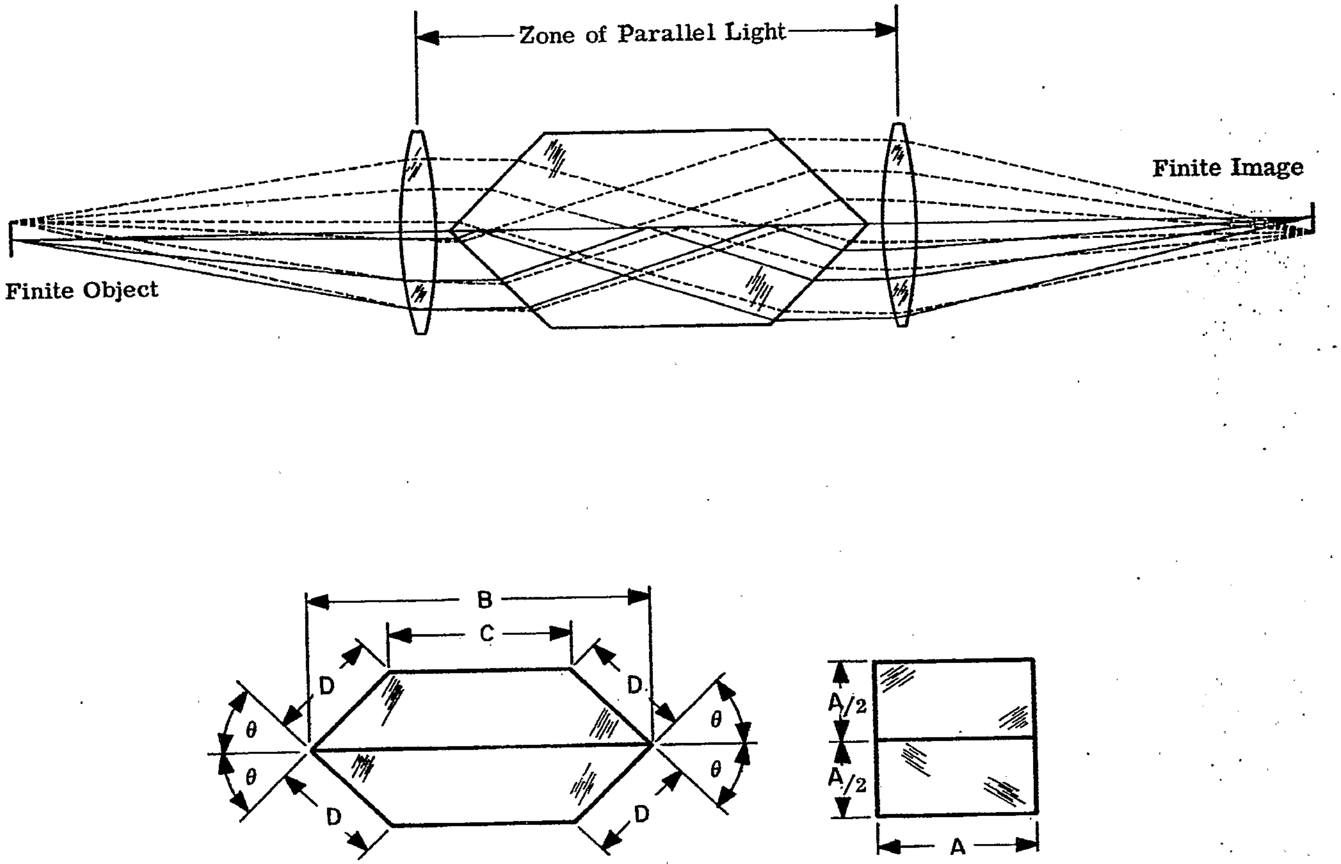


Figure 13.50-Dove prism tunnel diagram.

13.10.11 Double Dove Prism. This twin prism consists of two Harting-Dove prisms. Their reflecting surfaces are silvered and then the two halves are cemented together. This method cuts the length of the single Harting-Dove prism in half. This prism performs the duties of a single Harting-Dove prism, and it too must be placed in parallel light only.



$A = 1.00 \quad n = 1.5170 \quad C = B - A = 1.1136 \quad \theta = 45° \quad t = \dfrac{nA}{2 \sin\theta \sqrt{n^2 - \sin^2\theta} - \sin\theta} = nAC = 1.6893$

$t/n = 1.1135$

$$B = \frac{A}{2}\left[\frac{\sqrt{n^2 - \sin^2\theta} + \sin\theta}{\sqrt{n^2 - \sin^2\theta} - \sin\theta} + 1\right] = 2.1136A = 2.1136 \qquad D = \frac{A}{2\cos\theta} = 0.7071A = 0.7071$$

Figure 13.51-Double Dove prism.

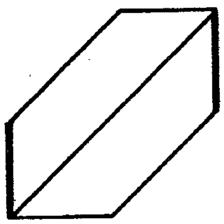
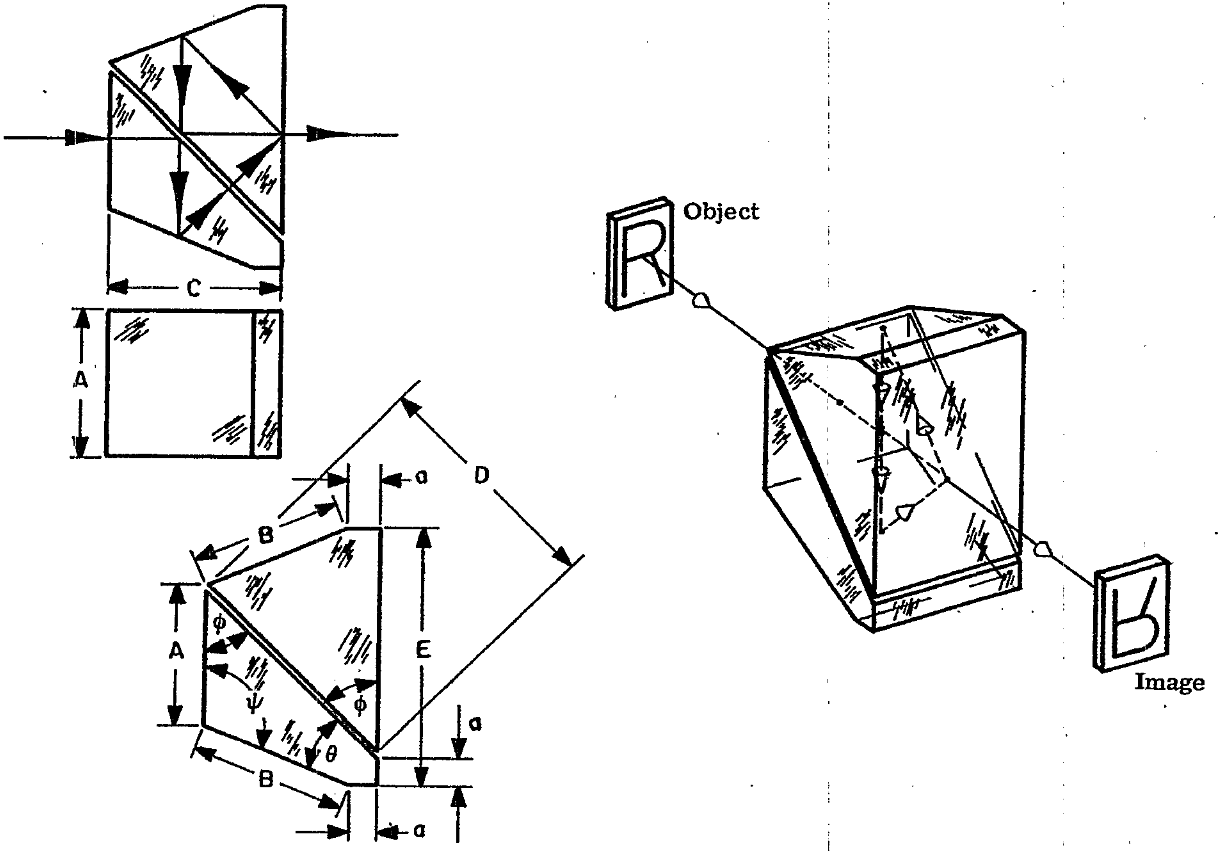


Figure 13.52-Double Dove prism tunnel diagram.

13.10.12 Pechan Prism. The prism performs the same duties as the Harting-Dove prism but it has one great advantage over the latter inasmuch as it may be placed in convergent or divergent light. This will permit the reduction in length or height of the instrument. It will invert (as shown) or revert the image, depending on its orientation. It may displace the line of sight if not properly centered but it will not deviate it. The surfaces marked B are silvered and covered with a protective coating. The unsilvered reflecting surfaces of the prism are separated by a distance of about 0.002 inch.



$A = 1.00$ $n = 1.5170$ $\theta = 22° 30'$ $\phi = 45°$ $\omega = 67° 30'$ $\psi = 112° 30'$ $a = 0.2071A = 0.2071$

$B = 1.0824A = 1.0824$ $C = 1.2071A = 1.2071$ $D = 1.7071A = 1.7071$ $E = 1.8284A = 1.8284$

$t = 4.6213A = 4.6213$ $t/n = 3.0464$

Figure 13.53-Pechan prism.

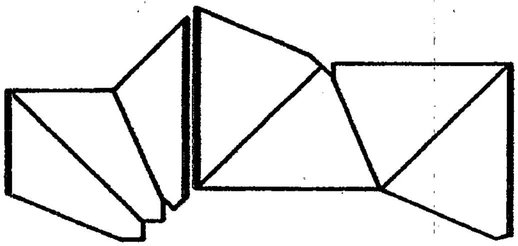
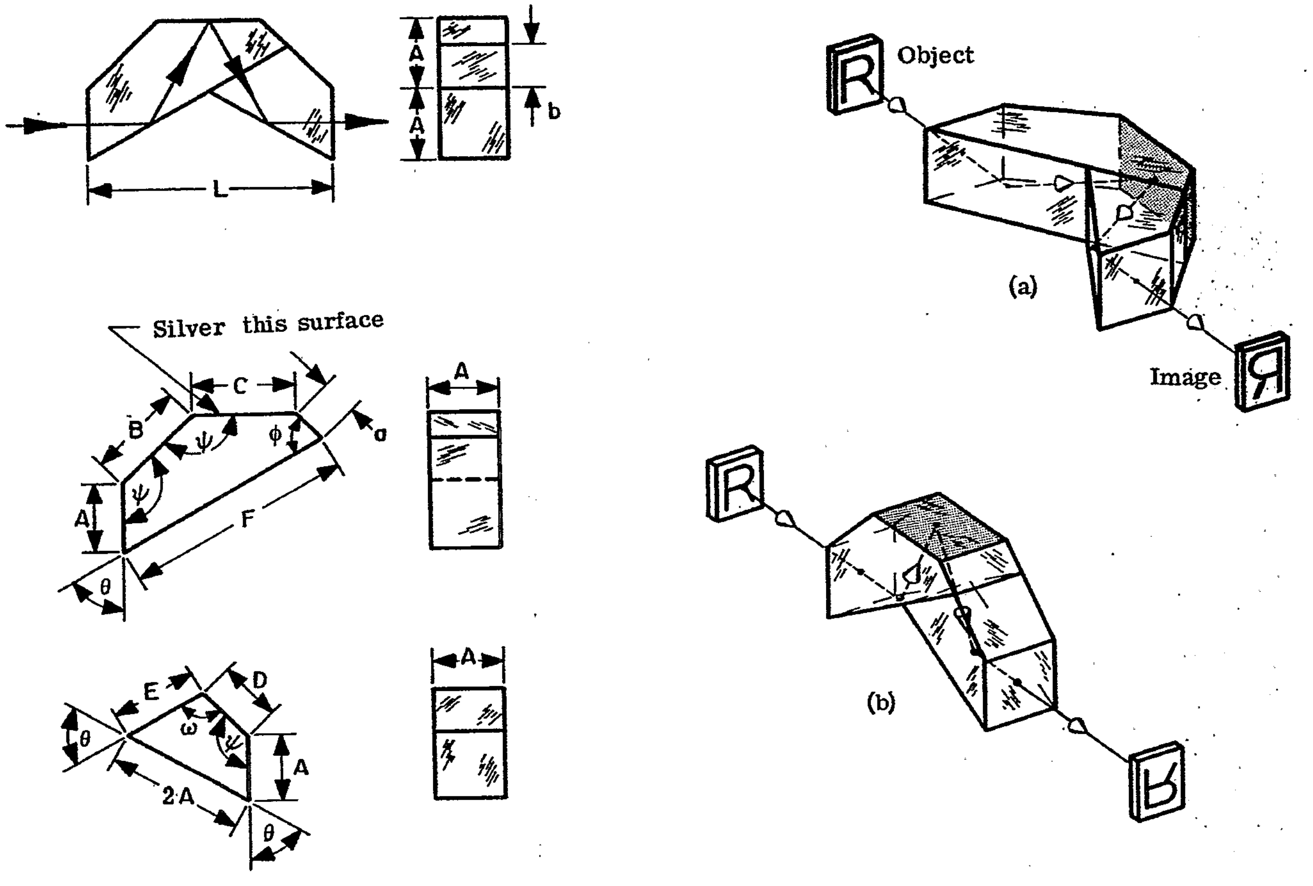


Figure 13.54-Pechan prism tunnel diagram.

13.10.13 Reversion Prism. This prism, which is a modification of the Abbe prism type A, consists of two elements which are cemented together. Like the Pechan prism, it may be placed in the path of parallel, converging, or diverging beams of light. Since three reflections are involved, it may be used to revert (a) or invert (b) the image, depending on its orientation. If not properly centered vertically, it will displace the line of sight by twice the centering error but will not deviate the sight line.



$A = 1.00$ $n = 1.5170$ $a = 0.5176A$ $b = 0.6340A$ $B = 1.4142A$ $C = 1.4641A$

$\theta = 60°$ $\phi = 75°$ $\psi = 135°$ $\omega = 105°$ $D = 0.8966A$ $E = 1.2679A$

$F = 3.2679A$ $L = 3.4641A$ $t/n = 3.4253A$ $t = 5.1962A$

Figure 13.55-Reversion prism.

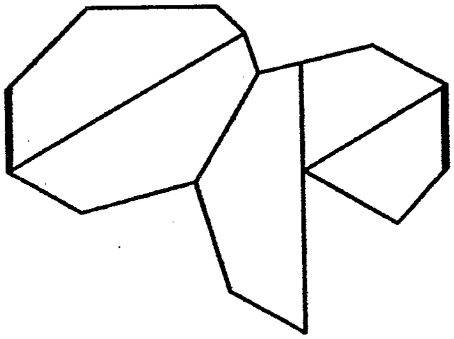
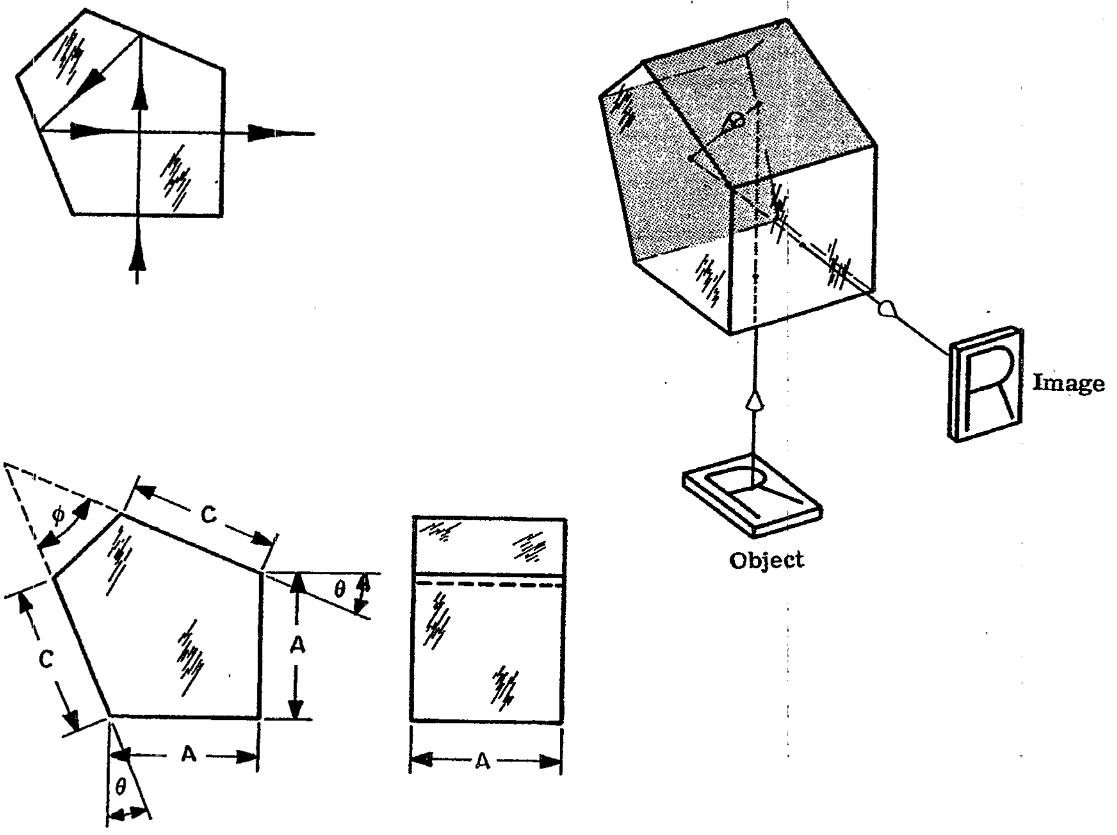


Figure 13.56-Reversion prism tunnel diagram.

13.10.14 Penta Prism. This prism will neither revert nor invert the image but will merely deviate the line of sight through an angle of 90°. The surfaces marked C in Figure 13.57 must be silvered and covered with a protective coating.



A = 1.00
n = 1.5170
θ = 22° 30'
φ = 45°
B = 0.4142 A = 0.4142
C = 1.0824 A = 1.0824
t = 3.4142 A = 3.4142
t/n = 2.2506

Figure 13.57-Penta prism.

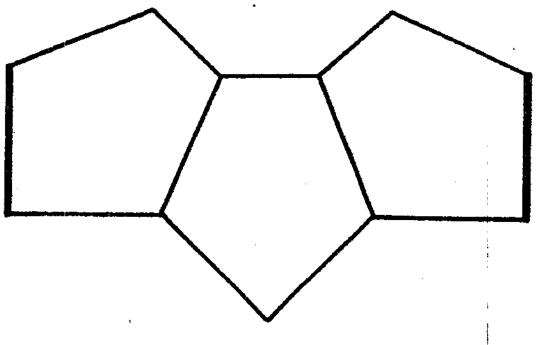
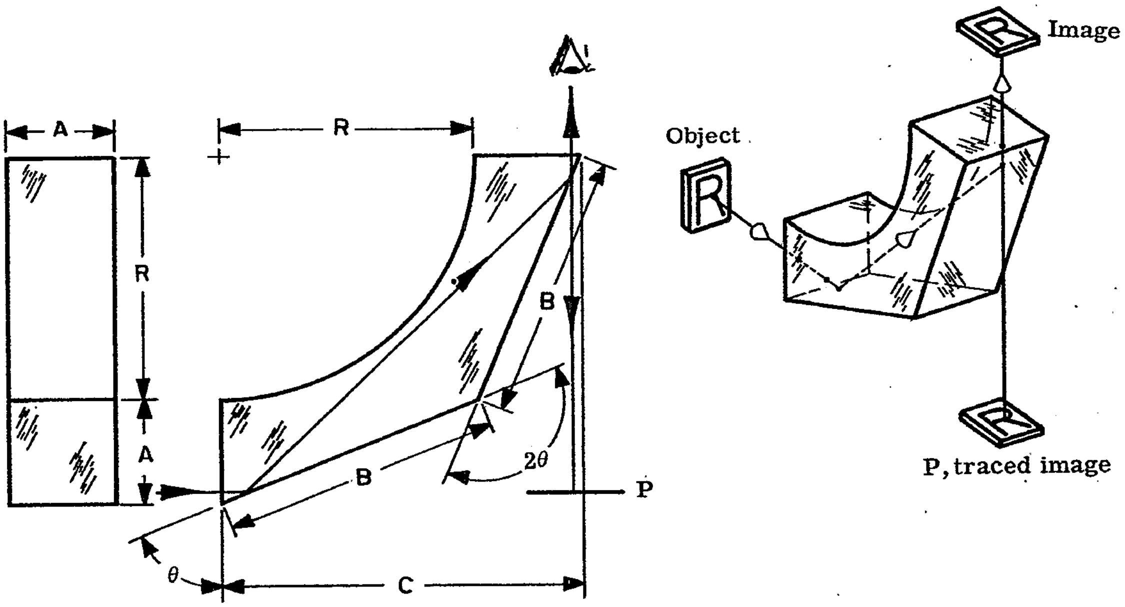


Figure 13.58-Penta prism tunnel diagram.

13.10.15 Wollaston Prism. Between the years of 1766 and 1828, the English scientist W. H. Wollaston designed a prism which has been named after him. It is made in one piece of glass and will neither invert nor revert the image, but it will deviate a beam of light through an angle of 90°. It is not used in military instruments due to its unfavorable shape. However, it is still used in an instrument known as "Camera Lucida," or "Camera Clara," the theory of which is explained here. If the observer's eye is placed right above the upper corner of the prism as shown in Figure 13.59, and a sheet of paper P is placed on the table about 10 inches from the eye, the observer will be able, with the aid of a pen, to trace the image of the object on the paper.



$A = 1.00$ $n = 1.5170$ $\theta = 67° 30'$ $B = 2.6131A = 2.6131$ $C = 3.4142A = 3.4142$

$R = 2.4142A = 2.4142$ $t = 2R = 4.8284A = 4.8284$ $t/n = 3.1829$

Figure 13.59-The Wollaston prism.

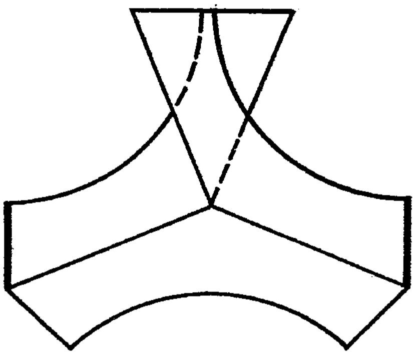


Figure 13.60-The Wollaston prism tunnel diagram.

13.10.16 Carl Zeiss Prism System. This combination consists of three single prisms (see Figure 13.61). As a rule the objective is placed between P_1 and P_2 ; however, it may also be placed in front of the objective prism P_1 . This system will invert and revert the image but will not deviate the line of sight. The line of sight will be displaced an amount depending on the distance between the prisms P_1 and P_2 .

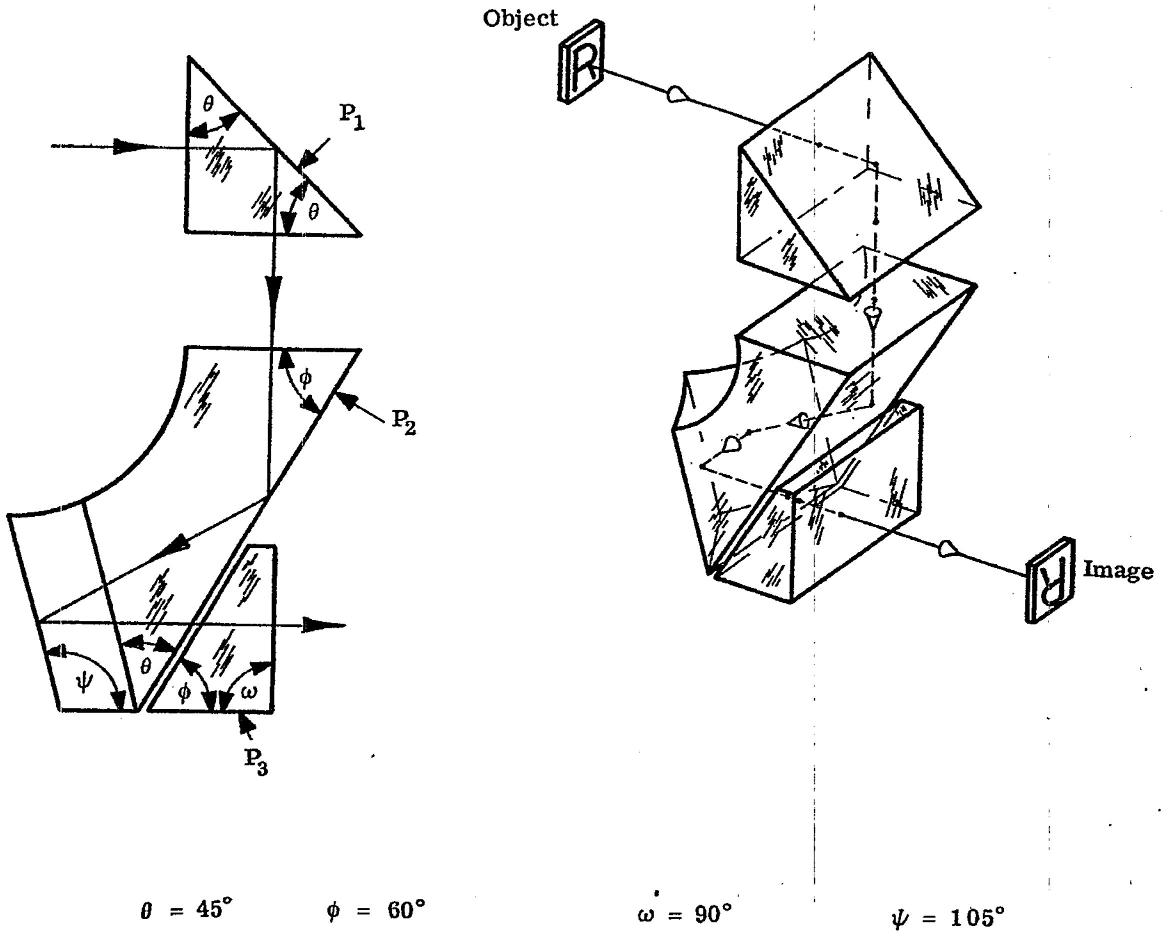


Figure 13.61-A Carl Zeiss prism system.

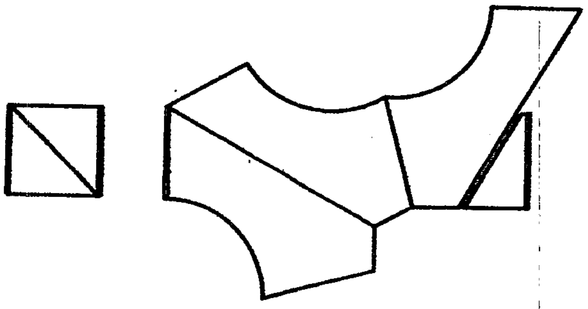
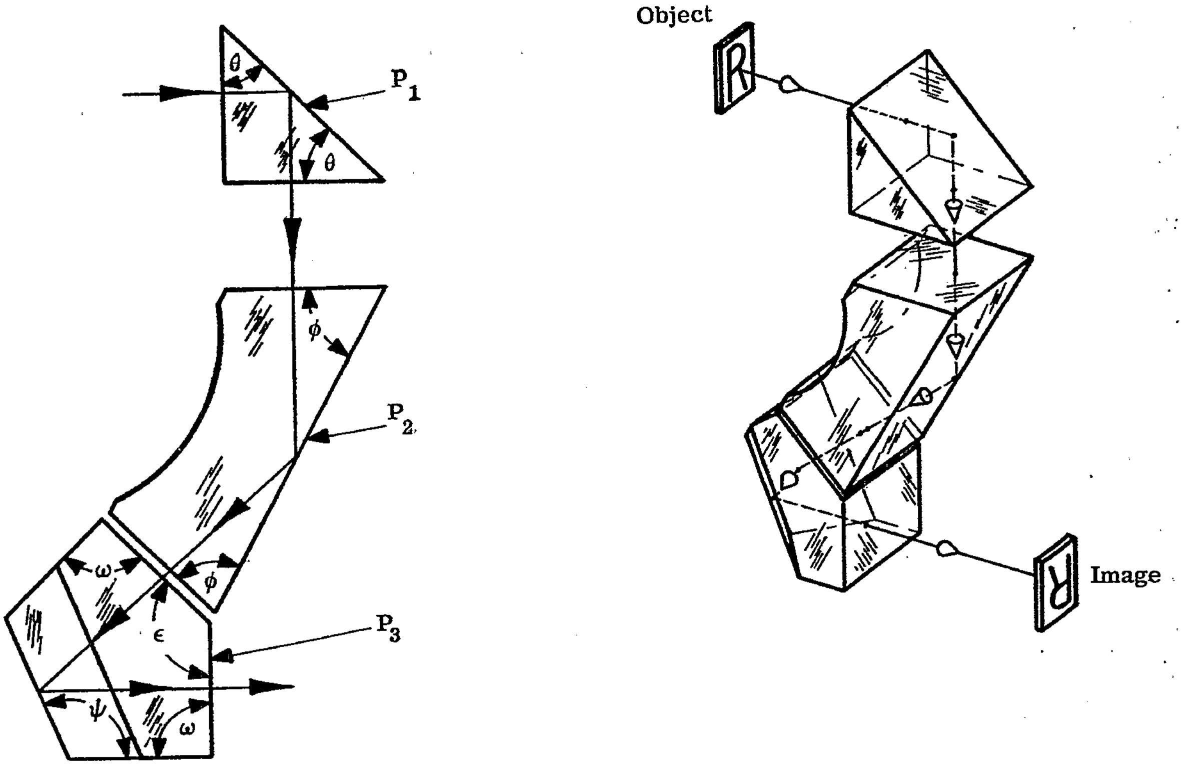


Figure 13.62-Carl Zeiss prism system tunnel diagram.

13.10.17 C. P. Goerz Prism System. This prism system consists of three single prisms as illustrated in Figure 13.63. The light is received by prism P_1, also known as the objective prism. The objective, usually placed between P_1 and P_2, may also be placed in front of P_1. This system will invert and revert the image. The line of sight will not be deviated from its original direction but will be displaced by an amount depending on the distance between the prisms P_1 and P_2.



$\theta = 45°$ $\phi = 67° 30'$ $\omega = 90°$ $\psi = 112° 30'$ $\epsilon = 135°$

Figure 13.63-A Goerz prism system.

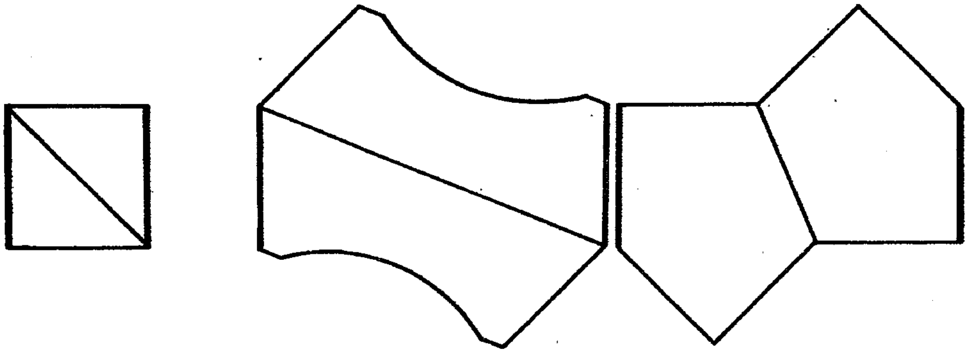


Figure 13.64-A Goerz prism system tunnel diagram.

13.10.18 Carl Zeiss Ocular Prism. This prism system, used in coincidence type range-finders, is made up of four single prisms, which are cemented together (see Figure 13.65). Light from the right will enter the system through the rhomboid prism P_1 and, after two internal reflections in this prism, and then three more in P_2 (the last reflection takes place on the silvered portion), the ray will emerge from the prism P_4 and then enter the eye of the observer. The image will be erect but reverted. Light from the left will enter the system through the prism P_3 and, after two internal reflections in this prism it will emerge from the system through P_4 and then it will also enter the eye of the observer. The image will appear inverted and reverted. In Figure 13.65 the refracting angles of the prism P_2, P_3, and P_4 are 22° 30'', and the light is deviated through an angle of 45°. This value may easily be varied by changing the refracting angles of the prisms.

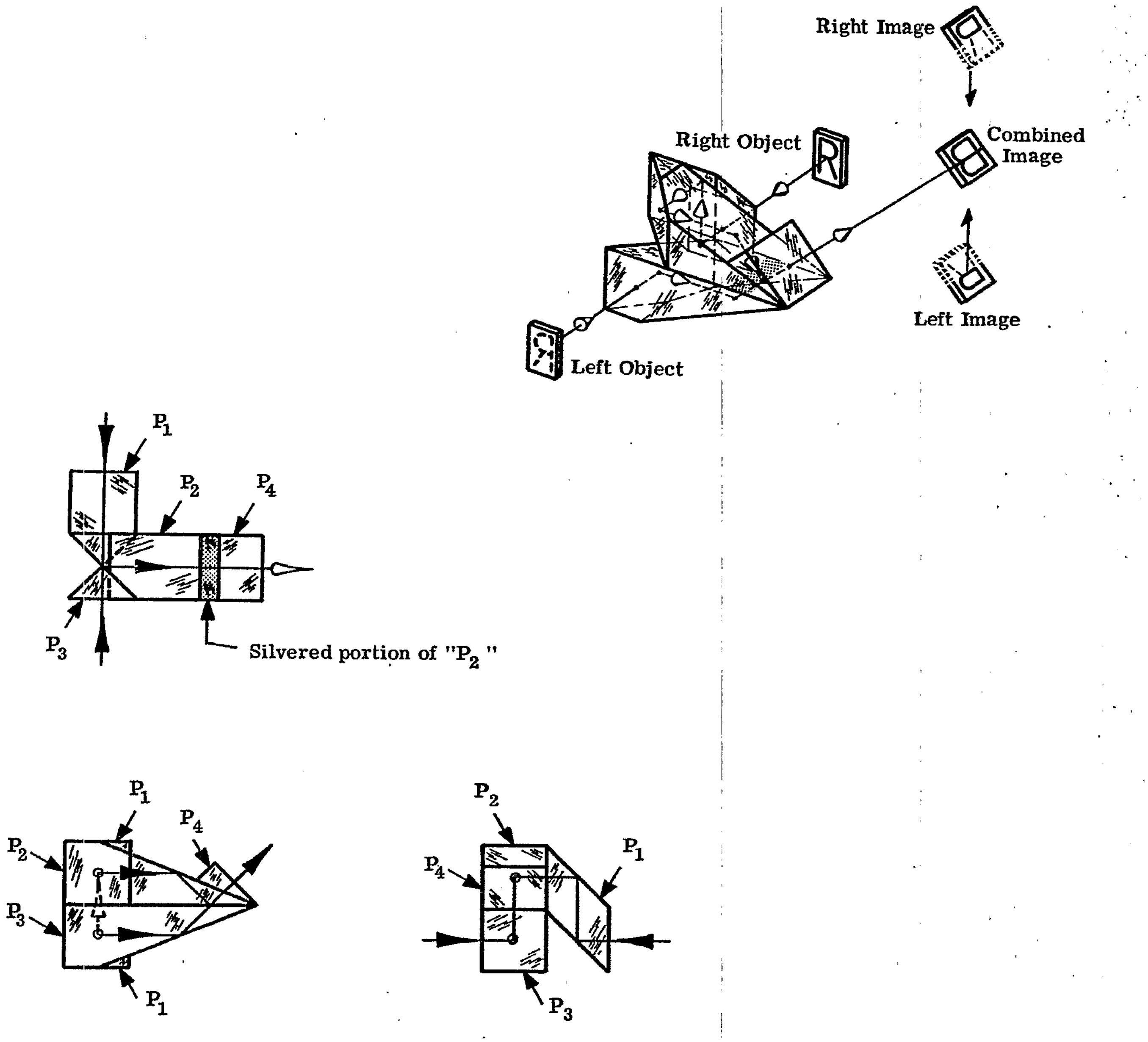


Figure 13.65-An ocular prism by Zeiss.

13.10.19 Barr and Stroud Ocular Prism. This ocular prism system, consisting of four single prisms and a cover, all cemented together, was used during the second world war by Research Enterprise Limited. It has one advantage over the Zeiss prism inasmuch as no silvered surface is required in producing the dividing line between the two images. On the other hand, the production division claims that the cost of manufacturing this prism is about five times that of the Zeiss prism, due to the great difficulties encountered in producing a well defined dividing line. The prisms P_1, P_2, and P_3 are made of a borosilicate crown glass ($n = 1.509$) and the prism P_4 of an extra dense flint glass ($n = 1.654$). The paths through the prism system of the various rays are illustrated in diagrams (a) and (b) of Figure 13.66. The rays of light, after passing through the right objective will enter the prism system through the prism P_1. After a reflection on the hypotenuse of this prism the rays will enter prism P_2, and, after three internal reflections in this prism, they will pass undeviated through the prism P_3 and the cover C and will then proceed towards the eyepiece. The image seen through this part of the prism system will appear inverted and reverted. The rays of light passing through the left objective will enter the prism system through the prism P_4, and will be reflected twice before they reach the dividing line between this prism and prism P_3. Due to the fact that the refractive index of P_4 is much greater than that of prism P_3, the rays will be reflected in an upward direction and emerge from the prism system parallel to the other rays. The image seen through this portion of the prism system will be erect but reverted.

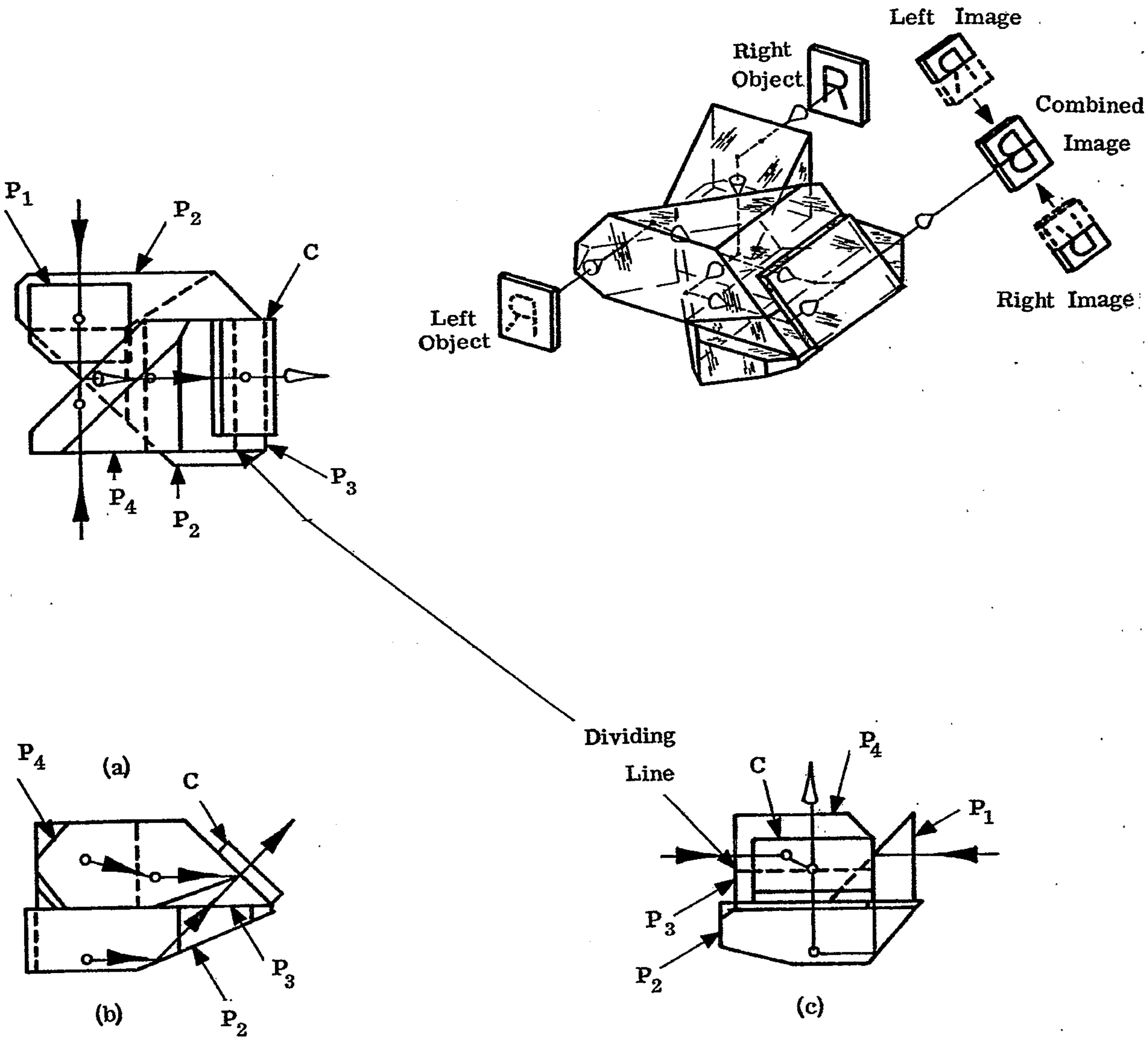
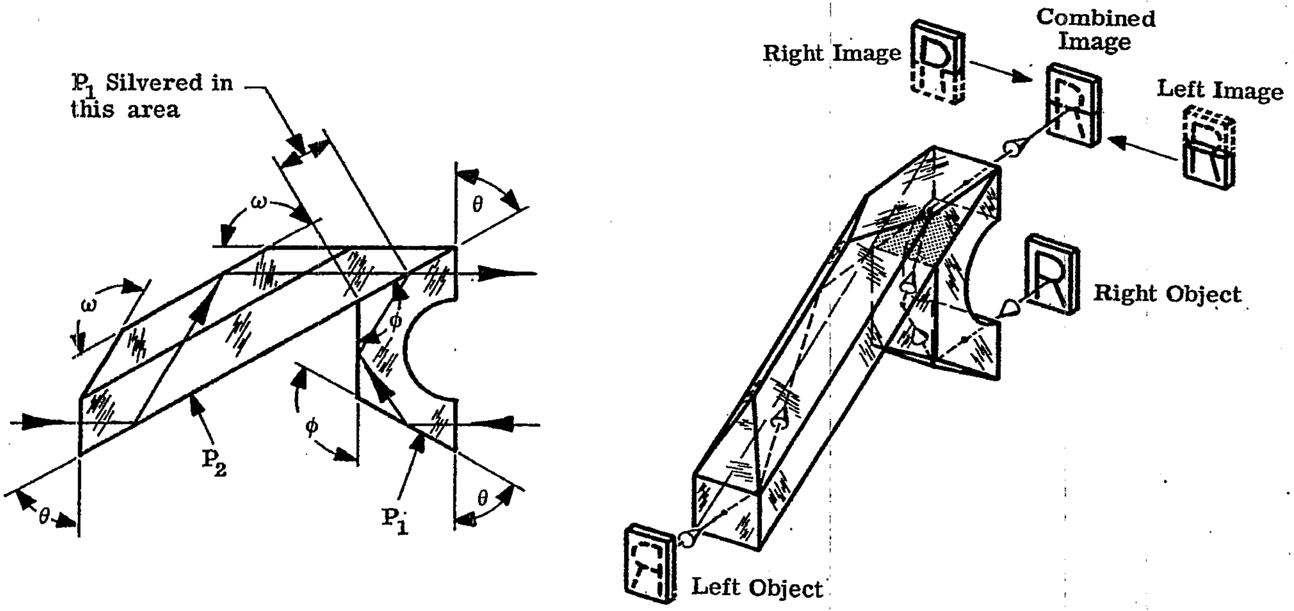


Figure 13.66-A Barr and Stroud ocular prism.

13.10.20 Carl Zeiss Coincidence Prism System. This prism system, illustrated in Figure 13.67, consists of two single prisms, P_1 and P_2. The lower half of the upper reflecting surface of P_1 is silvered and then the two prisms are cemented together. Light from the right will enter first P_1 at the lower entrance surface and, after three internal reflections it will emerge from the system at the upper exit surface. The image will appear reverted. Light from the left will enter through the prism P_2 and, after three reflections in the prism, it will enter prism P_1 through the unsilvered portion of the reflecting surface. The light will pass through P_1 undeviated before emerging from the prism system. The image will also appear reverted. This system is used in long base range finders. It is placed between the left objective and its image plane. The images formed by the left and right objective are formed in a plane normal to the line of sight through the dividing line of the silvered and the unsilvered portions of the reflecting surface. A lens erecting system will then transmit these images into the front focal plane of the ocular.



$\theta = 60°$ $\phi = 120°$ $\omega = 150°$

Figure 13.67-A Zeiss coincidence prism system.

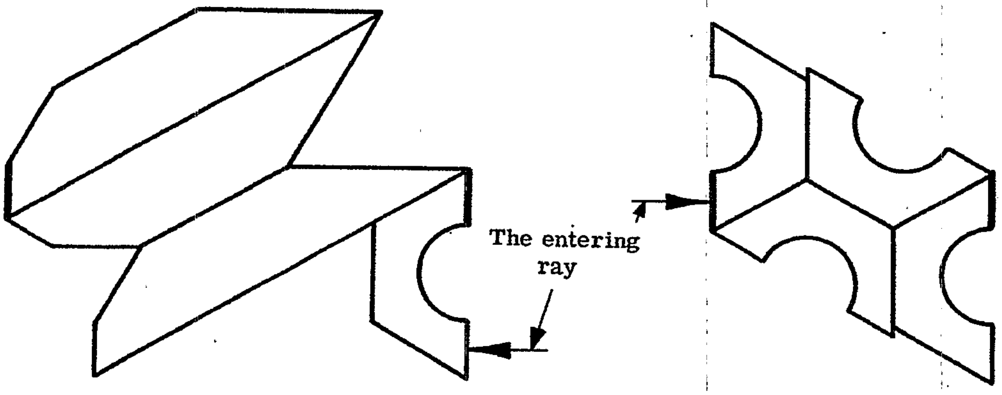
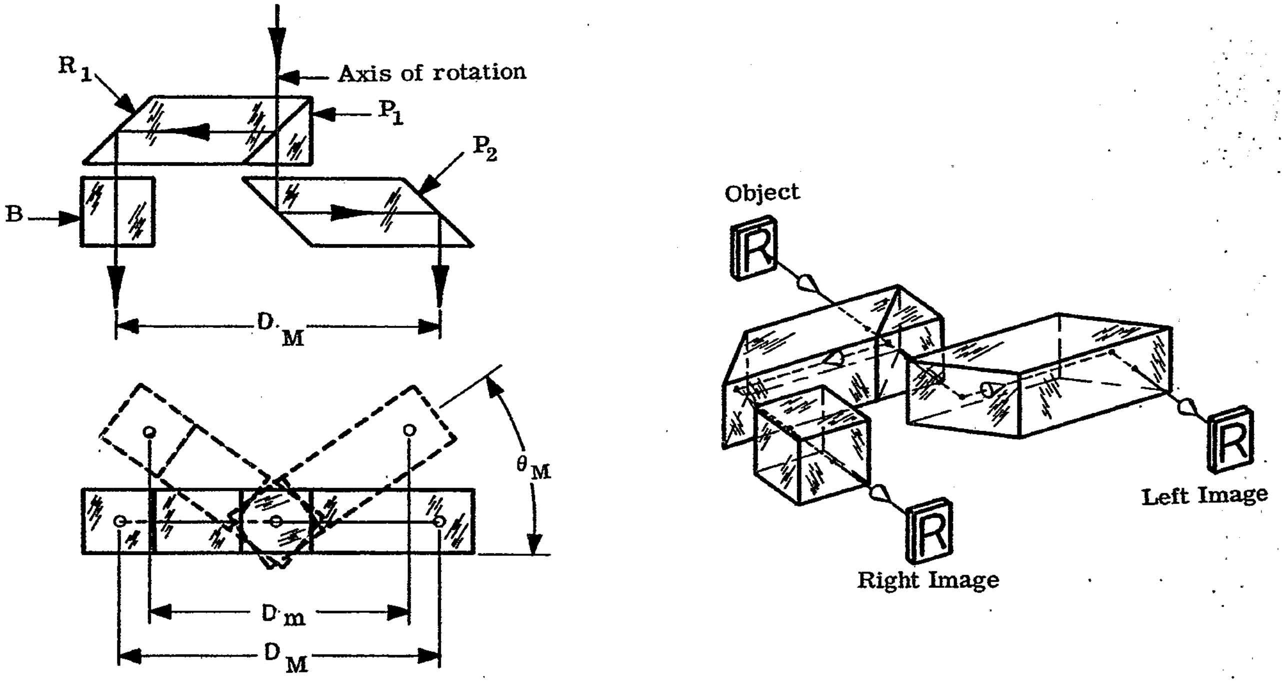


Figure 13.68-Zeiss coincidence prism tunnel diagram.

13.10.21 Carl Zeiss Binocular-Ocular Prism System. This system, illustrated in Figure 13.69, is used in binocular telescopes (or microscopes) when both eyes are to view the image presented by the objective. This system is made up of four single prisms, namely, the right angle prism P_1 cemented to the rhomboid prism R_1; the cemented surface will split the beam of light. The light passing through R_1 and P_1 will, before entering the eye, pass through the prism P_2. The other ray will pass through the block B which has been added to the system to equalize the length of the light-paths in glass. The interpupillary distance is designated by the letter D. Its value varies between the limits of D_m = 58 mm = 2.283 inches and D_M = 72 mm = 2.835 inches.



$$\cos \theta_M = \frac{D_m}{D_M} = \frac{2.283}{2.835} = 0.805291 \qquad \theta_M = 36^\circ\ 22'3''$$

Figure 13.69-A Carl Zeiss binocular-ocular prism system.

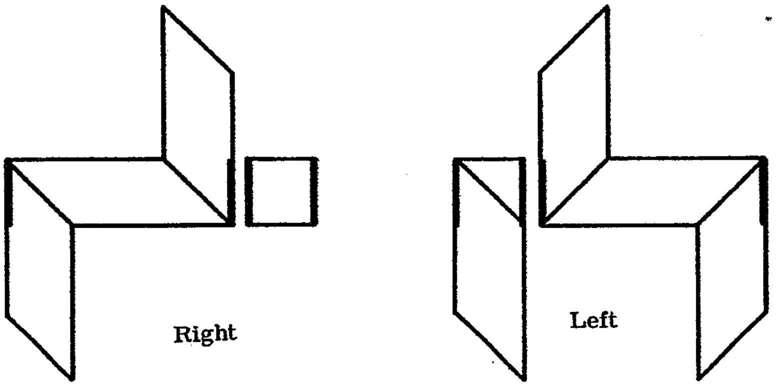
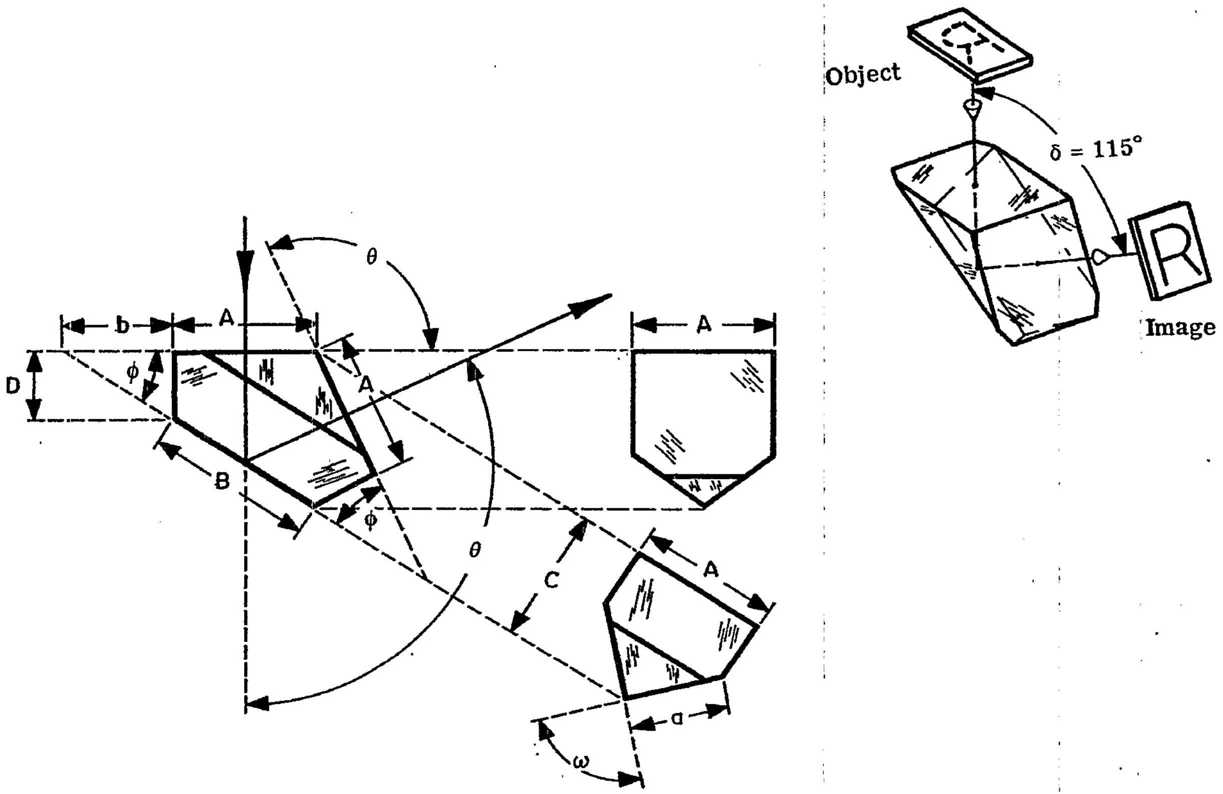


Figure 13.70-A Zeiss binocular-ocular prism tunnel diagram.

13.10.22 Frankford Arsenal Prism No. 1. This prism will revert the image and, at the same time, it will deviate the line of sight through an angle $\delta = 115°$.



$A = 1.00$ $n = 1.5170$ $\theta = 115°$ $\phi = 32° 30'$ $\omega = 90°$

$a = 0.7071A = 0.7071$ $b = 0.7320A = 0.7320$ $B = 1.1857A = 1.1857$ $C = 0.9306A = 0.9306$

$D = 0.4613A = 0.4613$ $t = 1.5697A = 1.5697$ $t/n = 1.0347$

Figure 13.71-Frankford Arsenal prism No. 1.

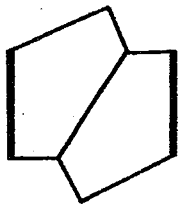
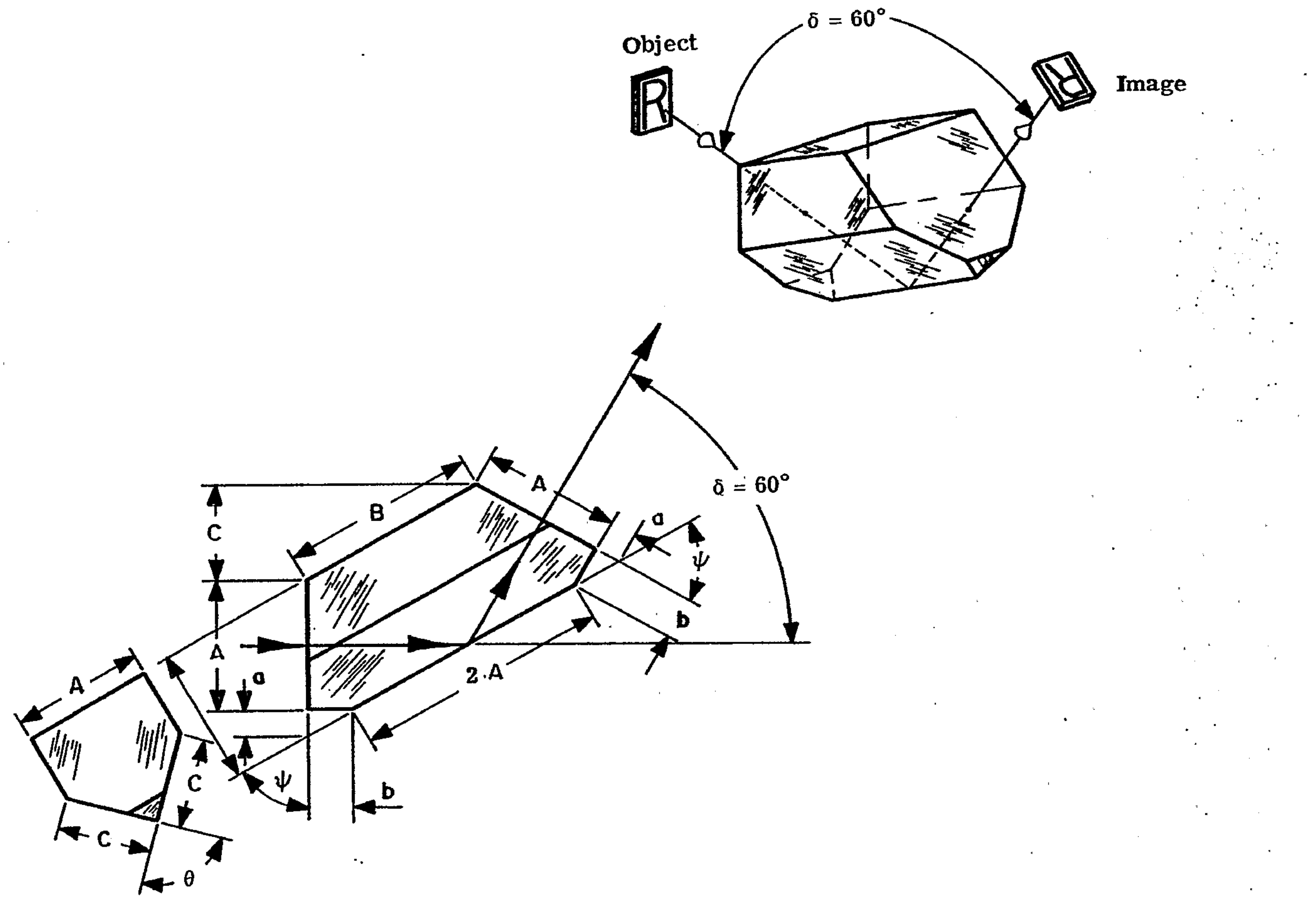


Figure 13.72-Frankford Arsenal prism No. 1 tunnel diagram.

13.10.23 Frankford Arsenal Prism No. 2. This prism is made in one piece. It will invert and revert the image and, at the same time, it will deviate the line of sight through an angle of $\delta = 60°$.



$A = 1.00$ $n = 1.5170$ $\theta = 90°$ $\psi = 60°$ $a = 0.1547A = 0.1547$ $b = 0.2680A = 0.2680$
$B = 1.4641A = 1.4641$ $C = 0.7321A = 0.7321$ $t = 2.2680A = 2.2680$ $t/n = 1.4951$

Figure 13.73-Frankford Arsenal prism No. 2.

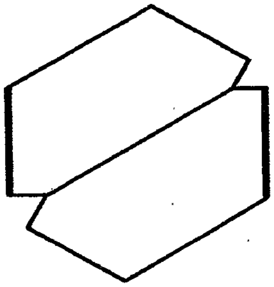
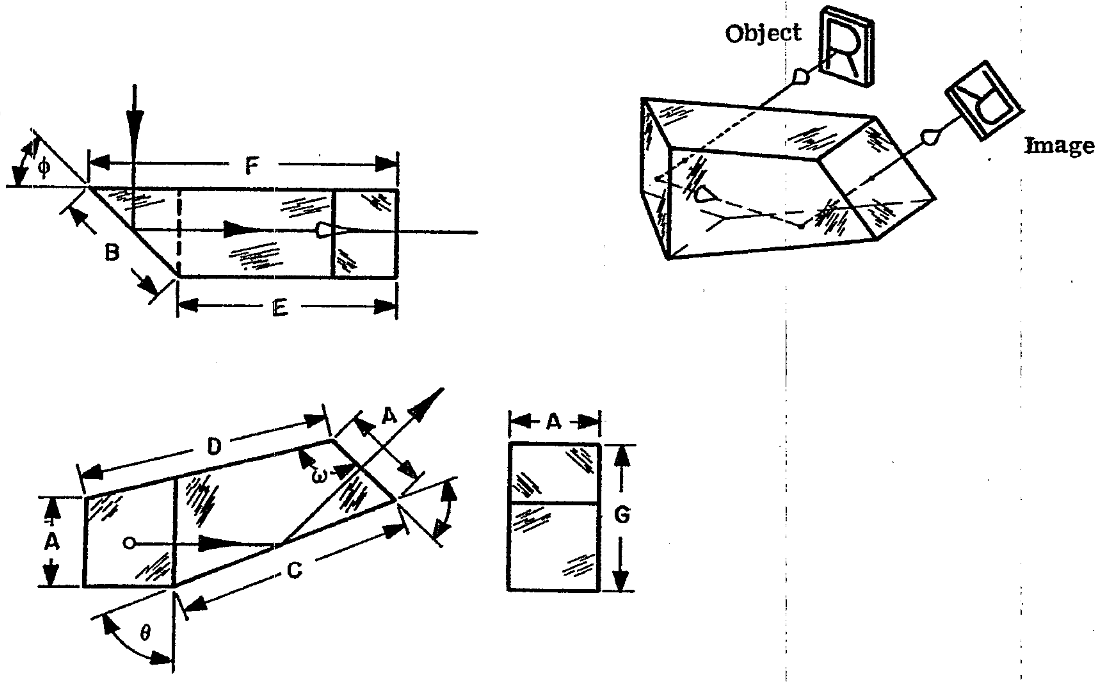


Figure 13.74-Frankford Arsenal prism No. 2 tunnel diagram.

13.10.24 Frankford Arsenal Prism No. 3. This prism is made in one piece. It will deviate the line of sight through an angle of 90° in the horizontal plane and, at the same time, through an angle of 45° in an upward direction. The observer, standing at right angles to the line of sight, will see an inverted and reverted image.



A = 1.00 n = 1.5170 θ = 67° 30' ϕ = 45° ω = 120° 21' 40" B = 1.4142A = 1.4142
C = 2.6131A = 2.6131 D = 2.7979A = 2.7979 E = 2.4142A = 2.4142 F = 3.4142A = 3.4142
G = 1.7071A = 1.7071 t = 3.4142A = 3.4142 t/n = 2.2506

Figure 13.75-Frankford Arsenal prism No. 3.

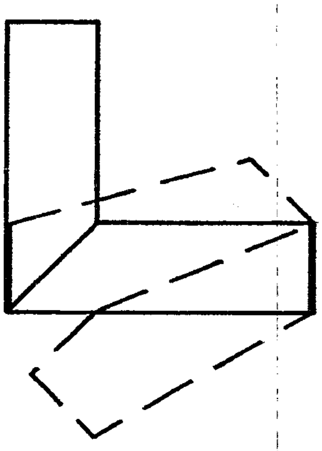
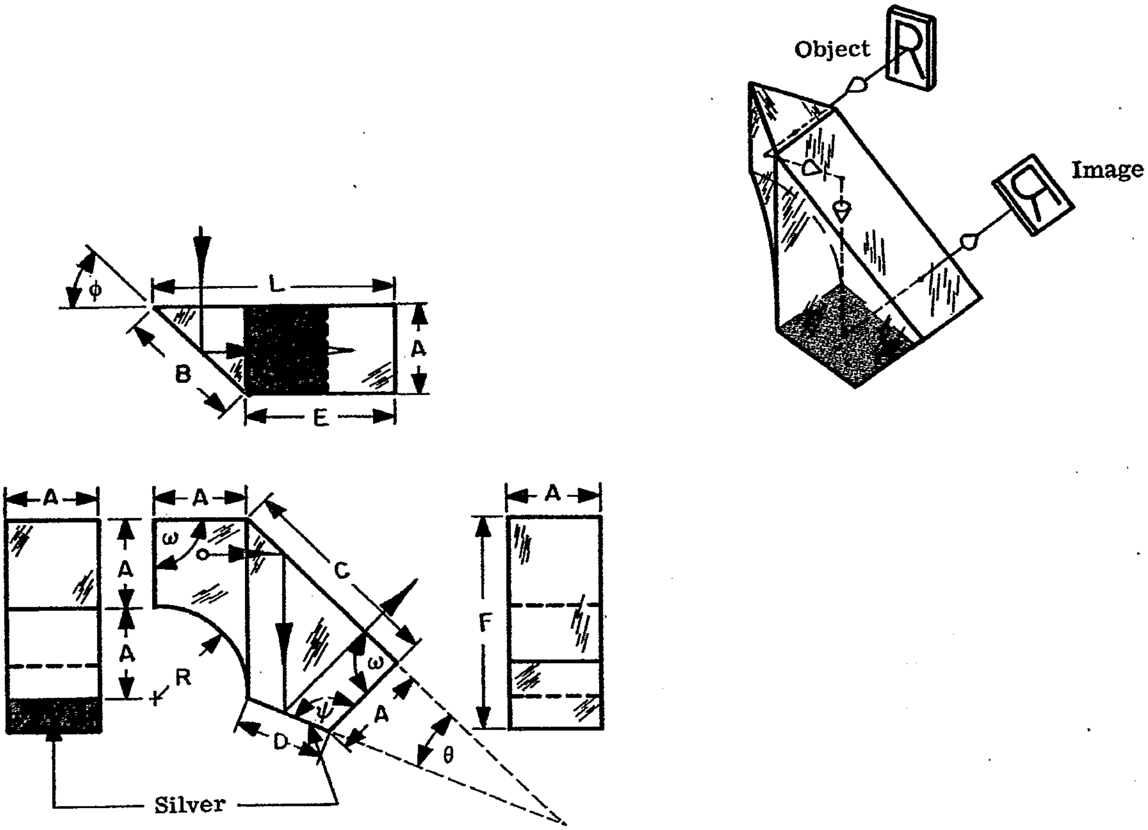


Figure 13.76-Frankford Arsenal prism No. 3 tunnel diagram.

13.10.25 Frankford Arsenal Prism No. 4. This prism is made of one piece of glass. The line of sight is deviated through an angle of 90° in the horizontal plane and, simultaneously, through an angle of 45° in the vertical plane. The observer, standing at right angles to the line of sight, will see the image reverted.



A = 1.00 n = 1.5170 θ = 22° 30' ϕ = 45° ω = 90° ψ = 112° 30' B = 1.4142A = 1.4142
C = 2.4142A = 2.4142 D = 1.0824A = 1.0824 E = 1.7071A = 1.7071 F = 2.4142A = 2.4142
L = 2.7071A = 2.7071 R = A = 1.00 t = 4.4142A = 4.4142 t/n = 2.9098

Figure 13.77-Frankford Arsenal prism No. 4.

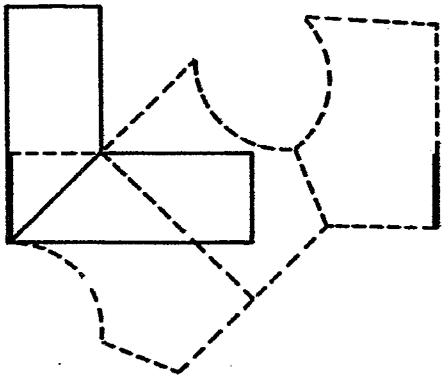
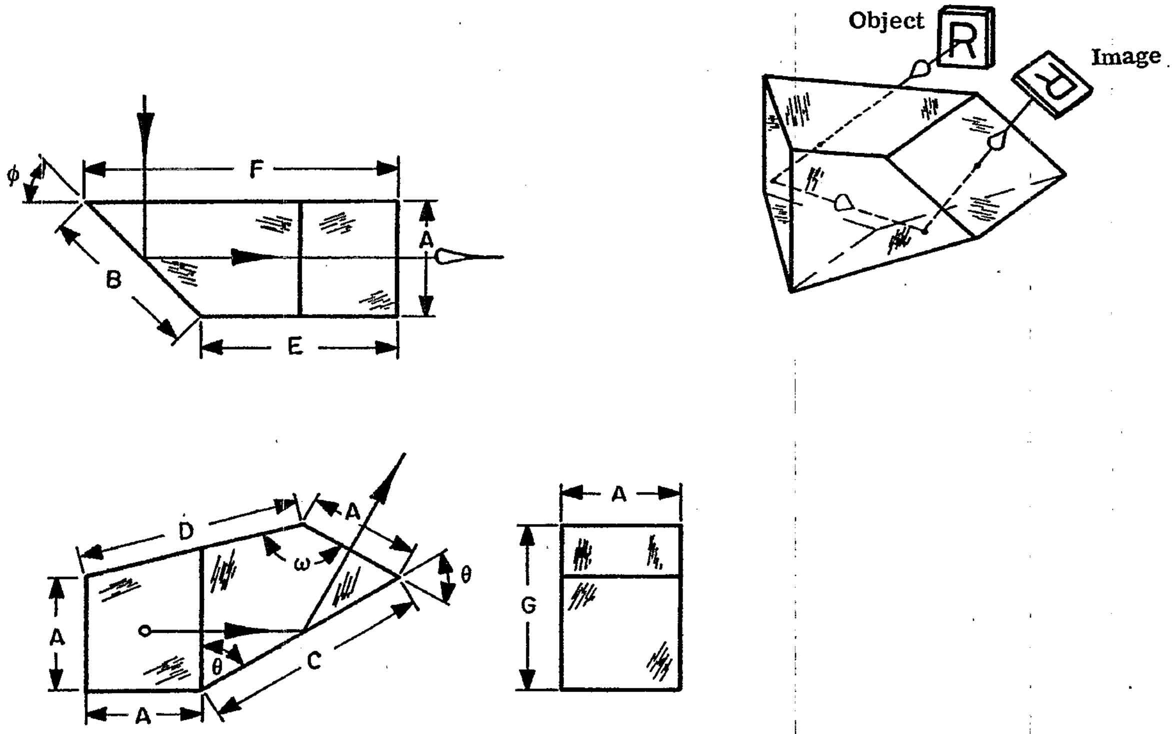


Figure 13.78-Frankford Arsenal prism No. 4 tunnel diagram.

13.10.26 Frankford Arsenal Prism No. 5. This prism is made in one piece. The line of sight is deviated through an angle of 90° in the horizontal plane and, simultaneously, through an angle of 60° in the vertical plane. The observer, standing at right angles to the line of sight, will see the image inverted and reverted.



$A = 1.00$ $n = 1.5170$ $\theta = 60°$ $\phi = 45°$ $\omega = 135°$ $B = 1.4142A = 1.4142$
$C = 2.000A = 2.000$ $D = 1.9318A = 1.9318$ $E = 1.7321A = 1.7321$ $F = 2.7321A = 2.7321$
$G = 1.500A = 1.500$ $t = 2.7437A = 2.7431$ $t/n = 1.8086$

Figure 13.79-Frankford Arsenal prism No. 5.

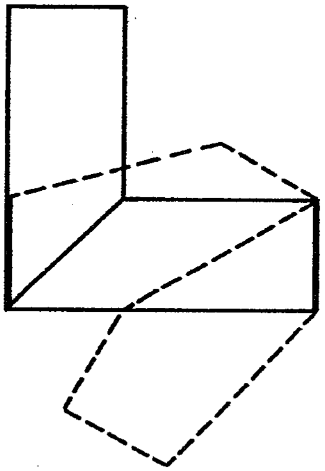
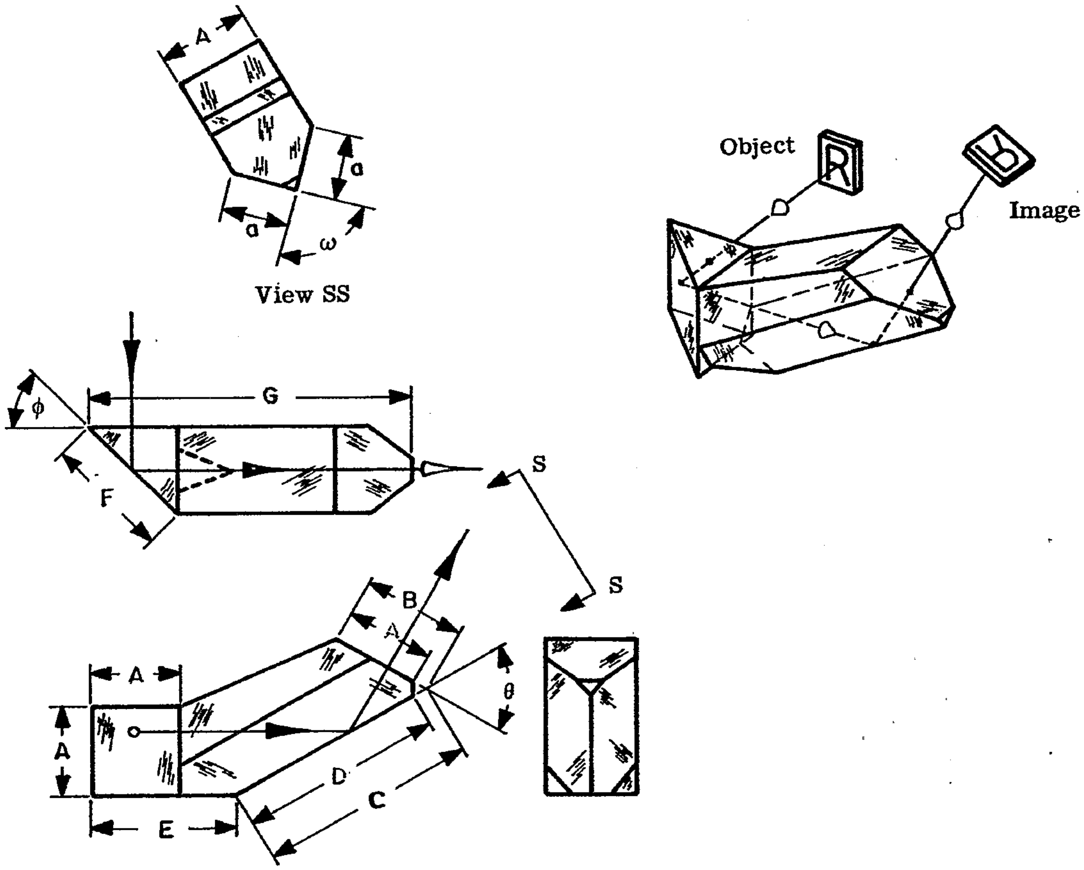


Figure 13.80-Frankford Arsenal prism No. 5 tunnel diagram.

13.10.27 Frankford Arsenal Prism No. 6. This prism is made in one piece. It will deviate the line of sight through an angle of 90° in the horizontal plane and through an angle of 60° in the vertical plane. The prism will invert the image.



A = 1.00 n = 1.5170 θ = 60° φ = 45° ω = 90° a = 0.7071A = 0.7071 t/n = 2.4180

B = 1.2071A = 1.2071 C = 2.4142A = 2.4142 D = 2.2071A = 2.2071 E = 1.5774A = 1.5774

F = 1.4142A = 1.4142 G = 3.4888A = 3.4888 H = 1.8107A = 1.8107 t = 3.6681A = 3.6681

Figure 13.81-Frankford Arsenal prism No. 6.

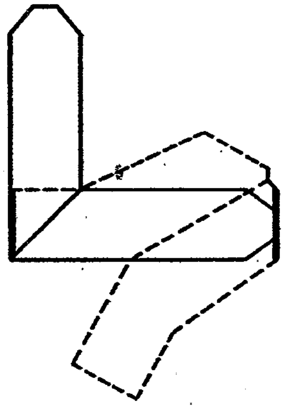
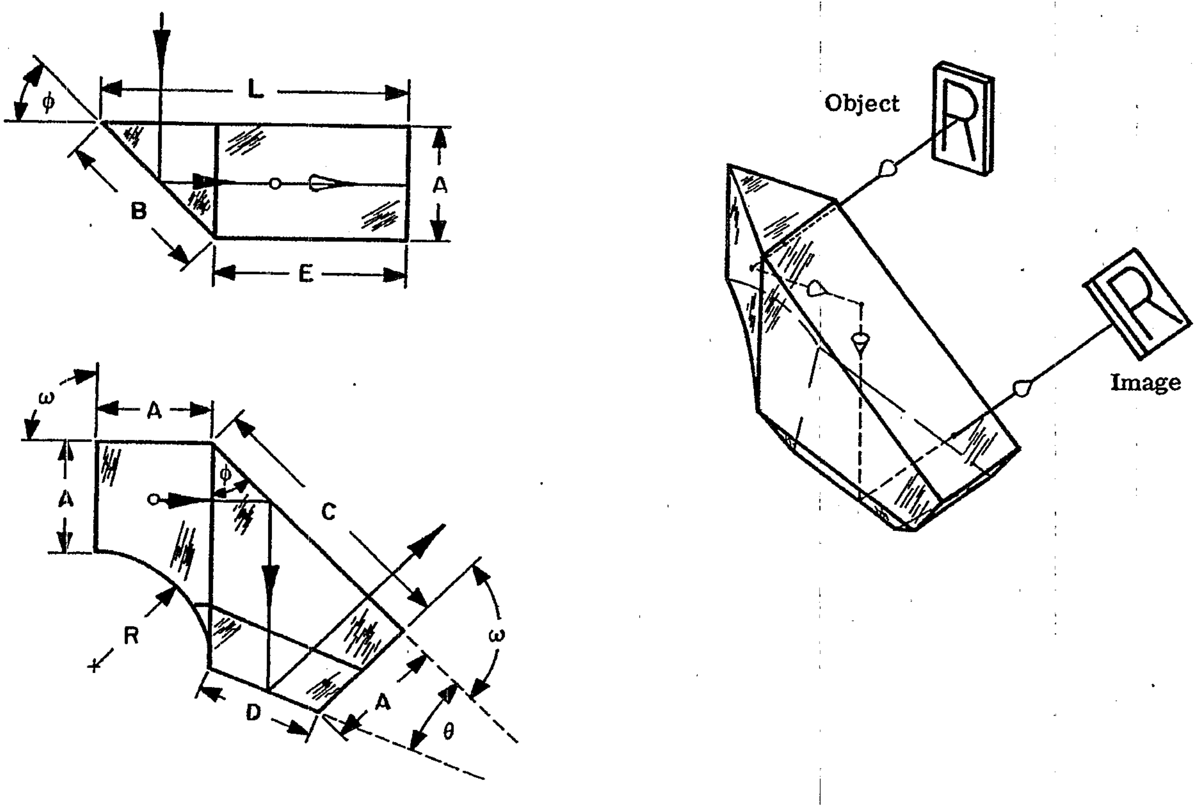


Figure 13.82-Frankford Arsenal prism No. 6 tunnel diagram.

13.10.28 Frankford Arsenal Prism No. 7. This prism is made in one piece. The line of sight is deviated through an angle of 90° in the horizontal plane and, simultaneously, through an angle of 45° in the vertical plane. The observer, standing at right angles to the line of sight, will see a normal image of the target since the prism neither inverts nor reverts the image.



$A = 1.00$ $n = 1.5170$ $\theta = 22° 30'$ $\omega = 90°$ $\phi = 45°$ $B = 1.4142A = 1.4142$ $R = A = 1.00$

$C = 2.4142A = 2.4142$ $D = 1.0824A = 1.0824$ $E = 1.7071A = 1.7071$ $L = 2.7071A = 2.7071$

$t = 4.4142A = 4.4142$ $t/n = 2.9098$

Figure 13.83-Frankford Arsenal prism No. 7.

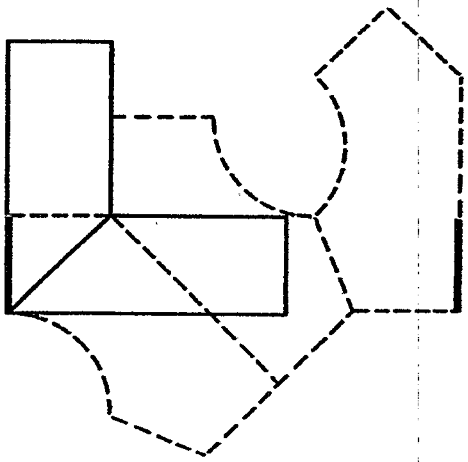


Figure 13.84-Frankford Arsenal prism No. 7 tunnel diagram.

14 EYEPIECES

14.1 GENERAL PRINCIPLES

14.1.1 Basic functions. The functions of the eyepiece were briefly described in Section 7.3. Let us now examine these more closely. The eyepiece in a visual instrument has three basic functions:

(1) It must, with the objective, form a good image of the object being viewed.

(2) It must serve as a magnifier if the instrument has a reticle.

(3) It must be designed so that the observer's eye can be placed in the exit pupil. Hence the exit pupil must be located at least 10 to 12 mm away from the last glass surface, this being the nearest the normal eye can approach the eyepiece surface with comfort.

14.1.2 Design considerations. Eyepieces should be designed to have a large apparent field of view (total field about 30° to 60°). Otherwise the viewer has the impression of looking down a tunnel towards a small opening. A large field of view necessitates bending the chief ray through an angle of $\alpha + \beta$, where α is the angle subtending one half the true field (field of view in object space), and β is the angle subtending one half the apparent field (field of view in image space). The chief ray must be bent with small spherical aberration so that the observer's eye may have a definite position in which to be located. Thus, one must design for a very large aperture lens with the aperture stop completely removed. Eyepieces are therefore very difficult to design. Very little can be done to improve the existing designs appreciably, nor is this a particularly fruitful area for a designer to spend time on. A more practical approach is to use or modify one of the existing designs. In the following paragraphs, several representative eyepieces are described which represent quite accurately the state of art in this field.

14.2 METHOD OF DESCRIPTION

14.2.1 General. In order to describe representative eyepieces on a comparative basis, the examples shown were designed for use with a ten power (10X or MP = 10) telescope. All eyepieces were designed to have a focal length of 2.54 cm and an exit pupil diameter of 5 mm. For each design a figure shows the shape of the lenses, and the location of the field stop and exit pupil. The eyepiece is shown with the exit pupil to the left and the objective is assumed to lie to the right. The reason for this representation is that it is generally easier to design a system with the object at infinity instead of the image at infinity. Hence eyepieces, as well as telescope objectives, are designed with the incident light assumed parallel. Similarly, microscope objectives as well as photographic objectives, are designed from long to short conjugate. The exit pupil is located by tracing a paraxial chief ray from the center of the objective (entrance pupil) back through the eyepiece. The exit pupil point is the intersection of the optical axis and this chief ray after it emerges from the system.

14.2.2 Descriptive details. In addition to the drawing, the following information is included.

(1) A table of curvatures, thickness, indices of refraction, ν-number, ΣP, and γ. This number, γ, is the ratio of the radius of the Petzval surface to the focal length of the eyepiece, and is used to estimate the field curvature in a complete telescope system. See Equation 11-(3).

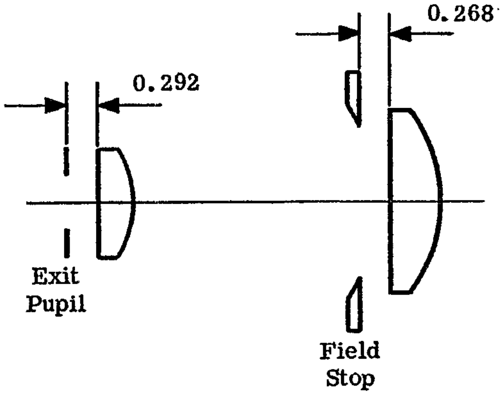
(2) Aberration curves in the focal plane of the eyepiece for parallel bundles of rays entering the eyepiece through the exit pupil. The curves are plotted in the same way as they were in Figures 8.7 and 8.8(b). The meridional fan and skew fan are shown on the same graph.

(3) The field curves, the distortion, and the lateral color. The first two are similar to the curves in Figures 8.10 and 8.9.

(4) A brief statement about each eyepiece.

14.3 THE HUYGENIAN EYEPIECE

14.3.1 Design data. Table 14.1 and Figures 14.1 through 14.3 present the design data and aberration curves for this eyepiece.



Scale - 1.35 to 1

Figure 14.1 - Huygenian eyepiece. Distances in cm.

c	t	Glass Type
0		
	0.3	517645
-0.9519		
	2.31	Air (n = 1)
0		
	0.45	517645
-0.6732		

$\Sigma P = -0.5538$
$\gamma = 0.711$

Table 14.1-Lens constants for the Huygenian eyepiece. Lengths in cm.

14.3.2 Use and characteristics. This eyepiece may be used where the apparent field of view is small (about ± 15°). It is commonly used in microscopes and small-field telescopes. The entire eyepiece is well corrected for lateral color. But because the field stop is located between the two lenses, and hence the field stop is viewed with the eyelens alone, its image will not be color free. Therefore, the Huygenian eyepiece is not recommended for use with a reticle in the field stop except in the special applications described in Section 23.3.5.2. It main virtue is its low cost. However, the eye relief (usually about 3 mm) afforded by this eyepiece is extremely short.

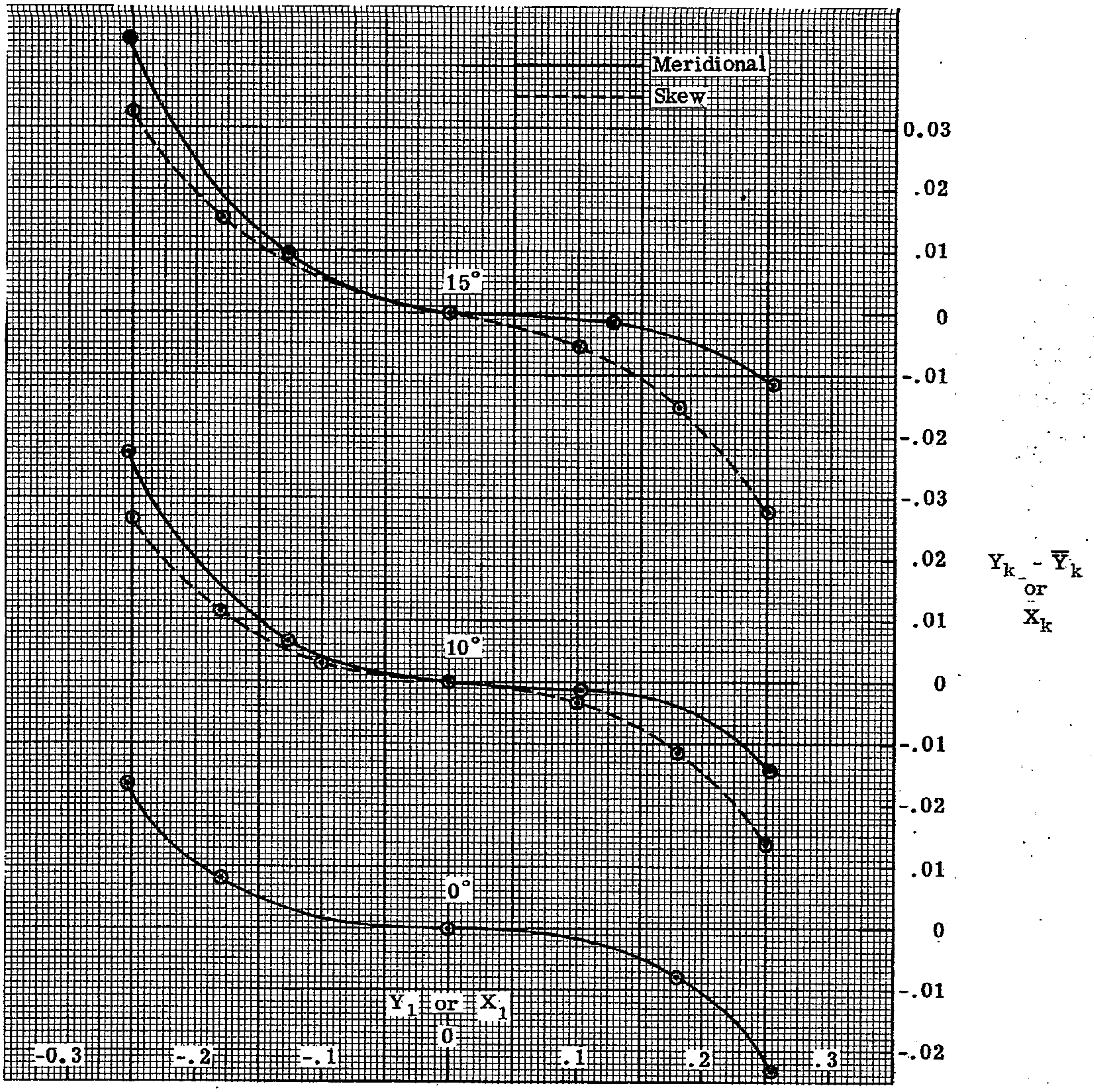


Figure 14.2- Meridional and skew fans for the Huygenian eyepiece.

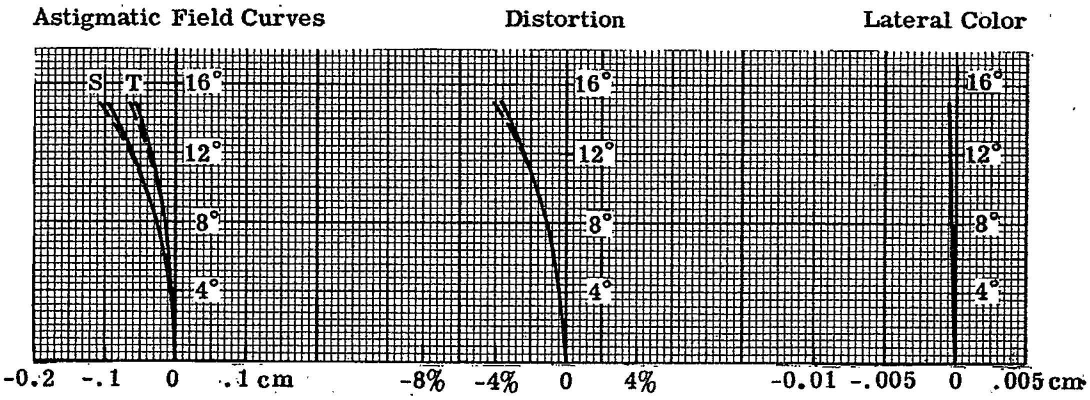
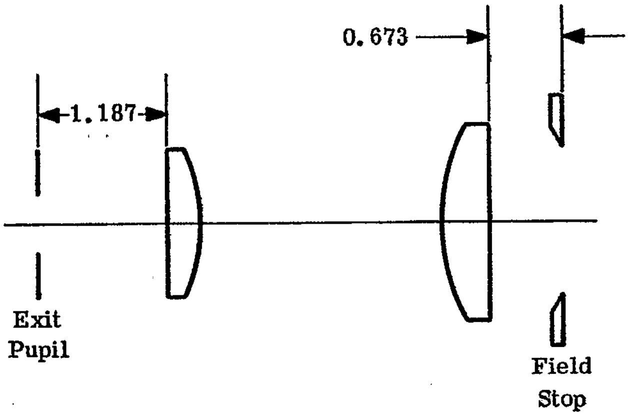


Figure 14.3 - Field, distortion, and lateral color curves for the Huygenian eyepiece.

14.4 THE RAMSDEN EYEPIECE

14.4.1 Design data. Table 14.2 and Figures 14.4 through 14.6 present the design data and aberration curves for this eyepiece.



Scale - 1.35 to 1

Figure 14.4 - Ramsden eyepiece. Distances in cm.

c	t	Glass Type
0		
	0.297	517645
-0.5712		
	2.116	Air (n = 1)
0.5077		
	0.424	517645
0		
	0.6733	
$\Sigma P = -0.3677$		
$\gamma = 1.07$		

Table 14.2 - Lens constants for the Ramsden eyepiece. Lengths in cm.

14.4.2 Use and characteristics. This eyepiece has a smaller field curve and is better corrected for the field stop plane than is the Huygenian. However, the lateral color is not corrected at all. At 15°, the lateral color is -0.007 cm which subtends an arc of 0.9 of a minute. This is well within tolerance, but if the field were extended beyond 15°, the color would become quite noticeable. The Ramsden is used in place of the Huygenian when cross hairs or reticles must be viewed. Like the Huygenian, its chief asset is its low cost. Its eye relief is still short (about 12 mm), but better than that of the Huygenian.

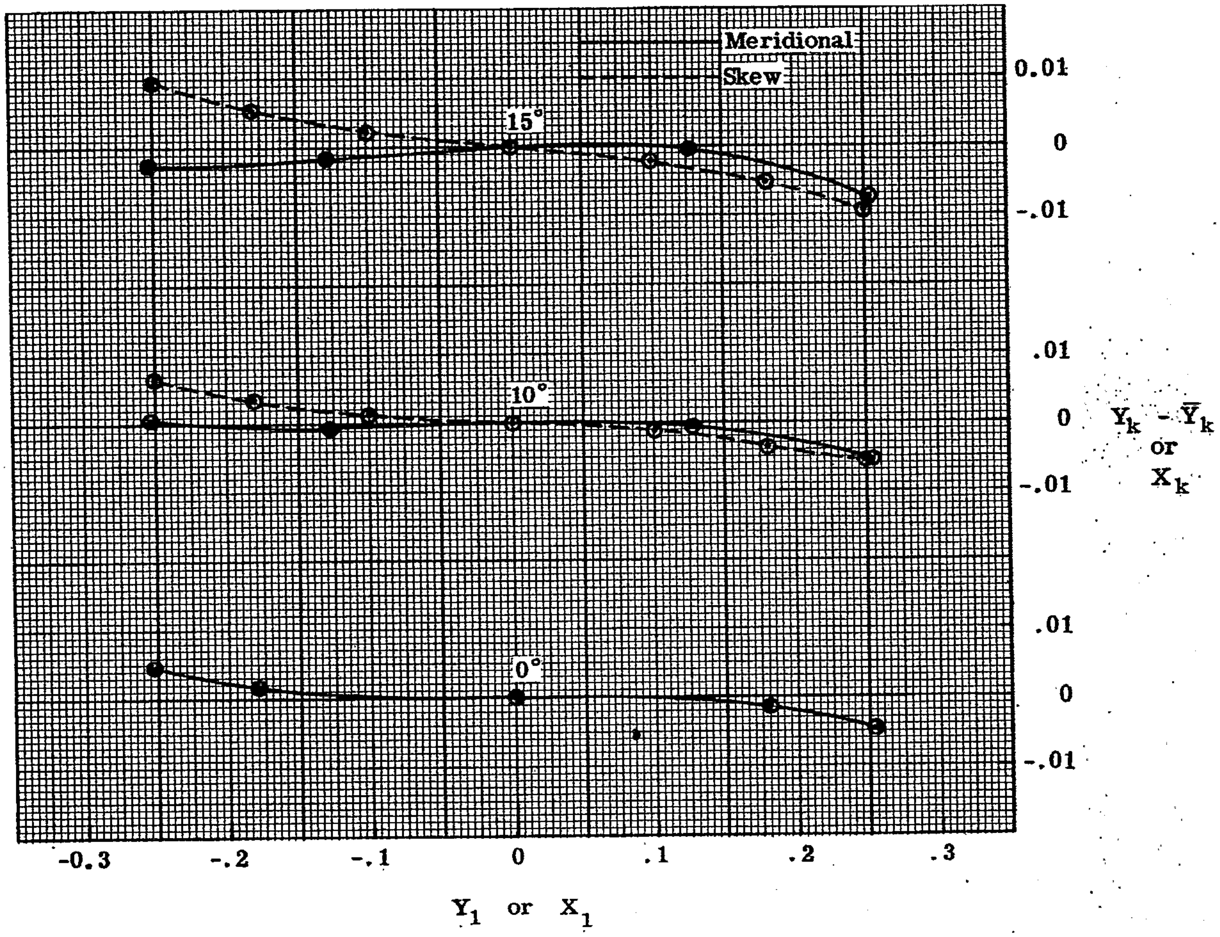


Figure 14.5- Meridional and skew fans for the Ramsden eyepiece.

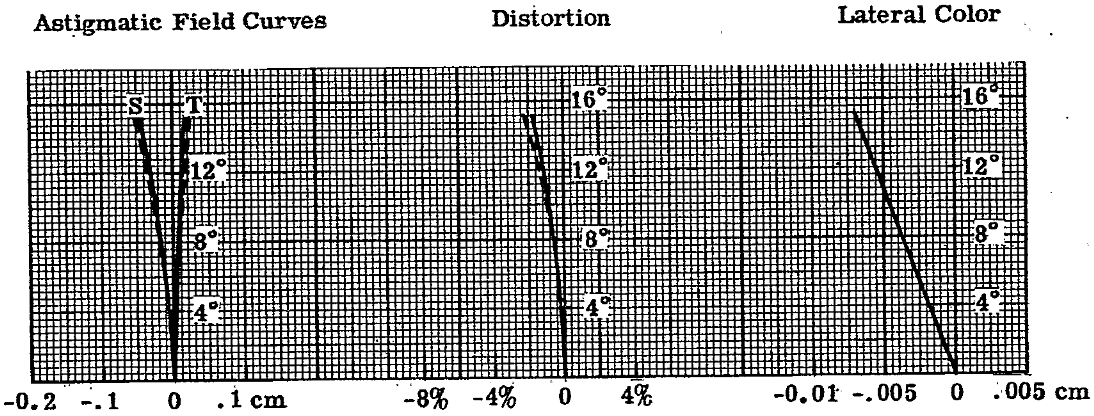
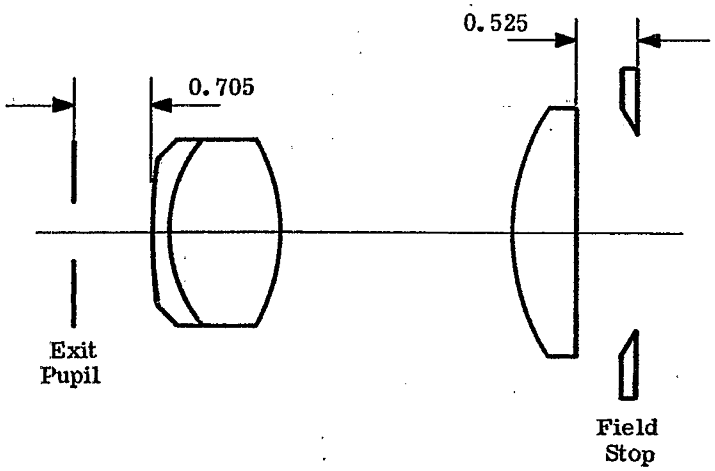


Figure 14.6- Field, distortion, and lateral color curves for the Ramsden eyepiece.

14.5 THE KELLNER EYEPIECE

14.5.1 Design data. Table 14.3 and Figures 14.7 through 14.9 present the design data and aberration curves for this eyepiece.



Scale - 1.35 to 1

Figure 14.7 - Kellner eyepiece. Distances in cm.

c	t	Glass Type
0.1039		
	0.159	617366
0.7393		
	0.995	541599
-0.5525		
	2.089	Air (n = 1)
0.4699		
	0.577	541599
0		
	0.5251	
$\Sigma P = -0.3760$		
$\gamma = 1.047$		

Table 14.3 - Lens constants for the Kellner eyepiece. Lengths in cm.

14.5.2 Use and characteristics. The Kellner eyepiece is partially corrected for lateral color and is used out to 20° half angle. It is probably the most common eyepiece used in moderately wide field telescopic systems. The eye relief (about 7 mm) is intermediate between the Huygenian and as the Ramsden eyepieces.

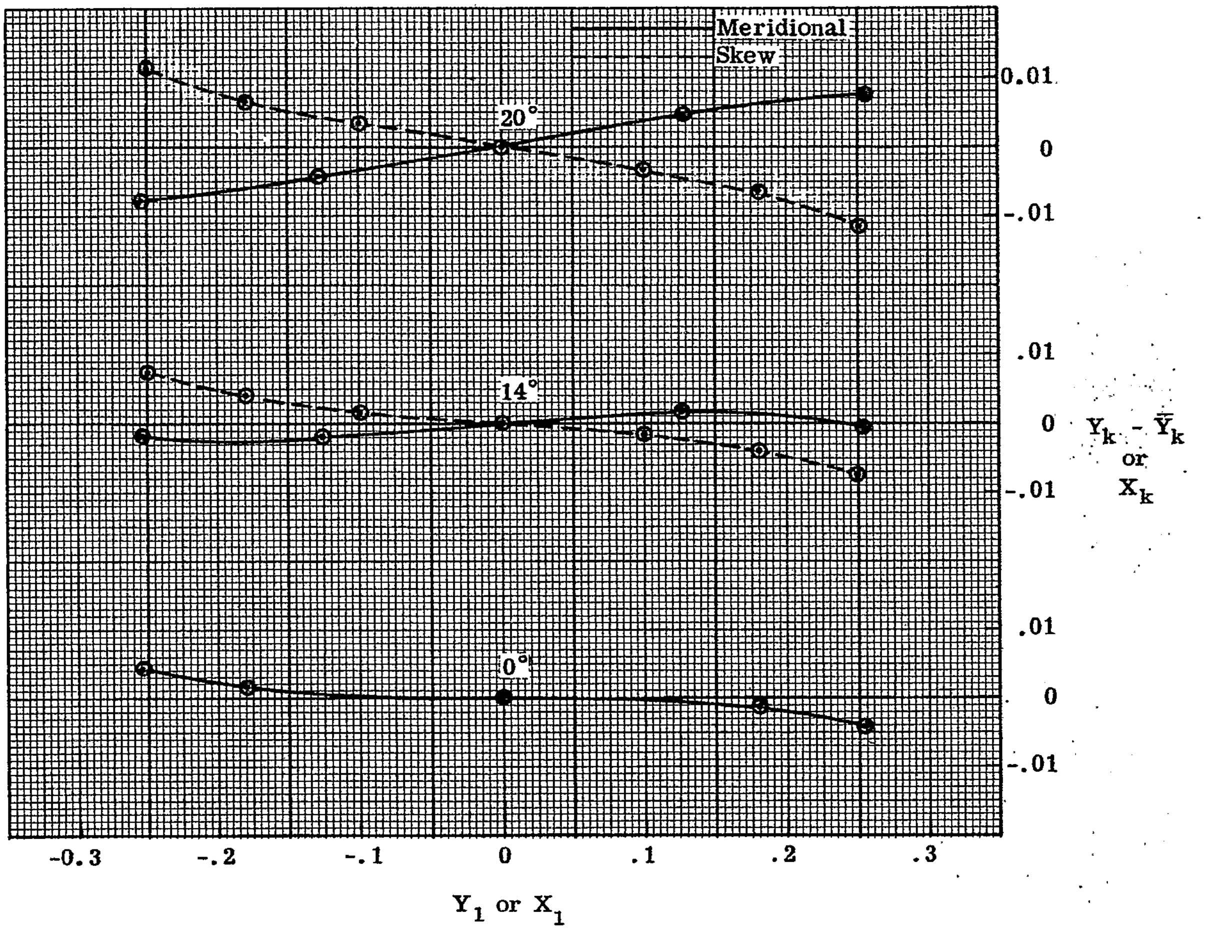


Figure 14.8- Meridional and skew fans for the Kellner eyepiece.

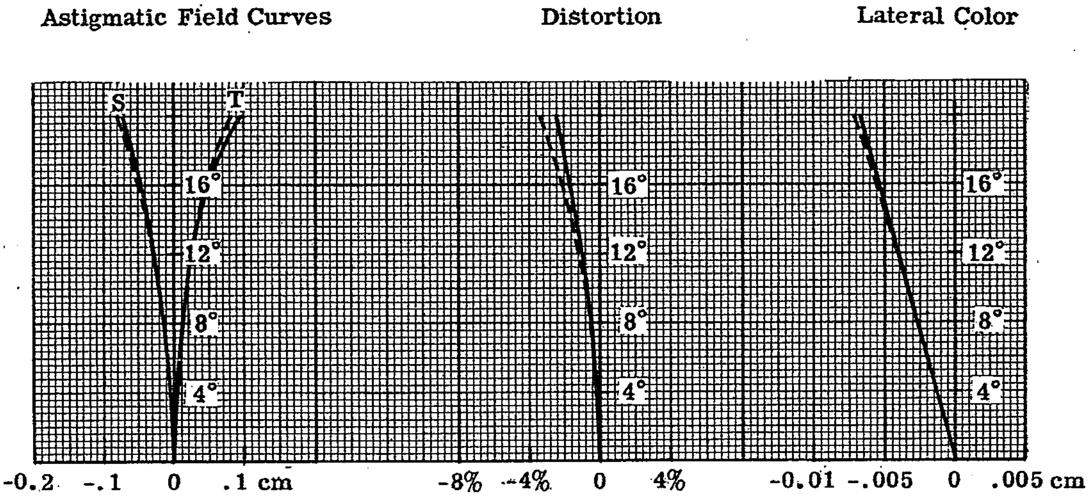


Figure 14.9- Field, distortion, and lateral color curves for the Kellner eyepiece.

14.6 THE ORTHOSCOPIC EYEPIECE

14.6.1 Design data. Table 14.4 and Figures 14.10 through 14.12 present the design data and aberration curves for this eyepiece.

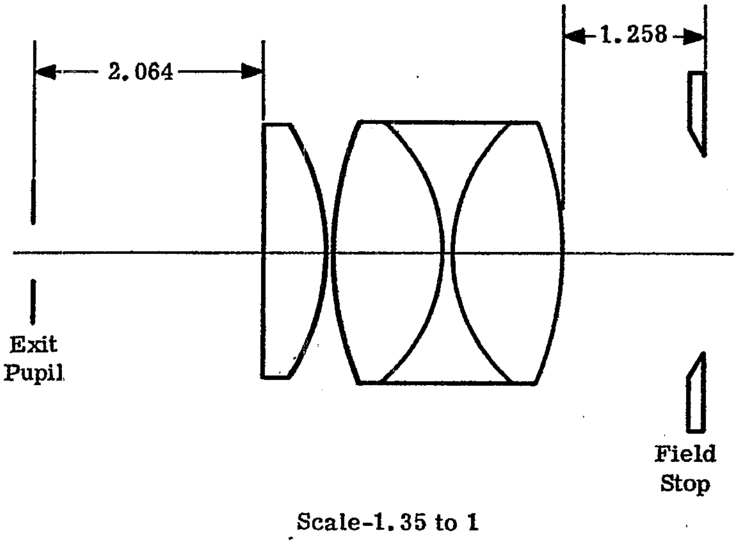


Figure 14.10 - Orthoscopic eyepiece. Distances in cm.

c	t	Glass Type
0		
	0.582	573574
-0.4398		
	0.0276	Air (n = 1)
0.3089		
	0.9921	513605
-0.6281		
	0.1012	617366
0.6281		
	0.9921	513605
-0.3089		
	1.258	

$\Sigma P = -0.3158$

$\gamma = 1.25$

Table 14.4 - Lens constants for the orthoscopic eyepiece. Lengths in cm.

14.6.2 Use and characteristics. The orthoscopic eyepiece has several advantages: (a) the γ is larger than for the two previous examples, hence the Petzval curvature is smaller; (b) the lateral color is very well corrected; and (c) it has a long eye relief (about 20 mm). However the T field has a tendency to fly backward rapidly. In more expensive instruments this eyepiece is used instead of the Kellner, sometimes as far out as 25° half angle.

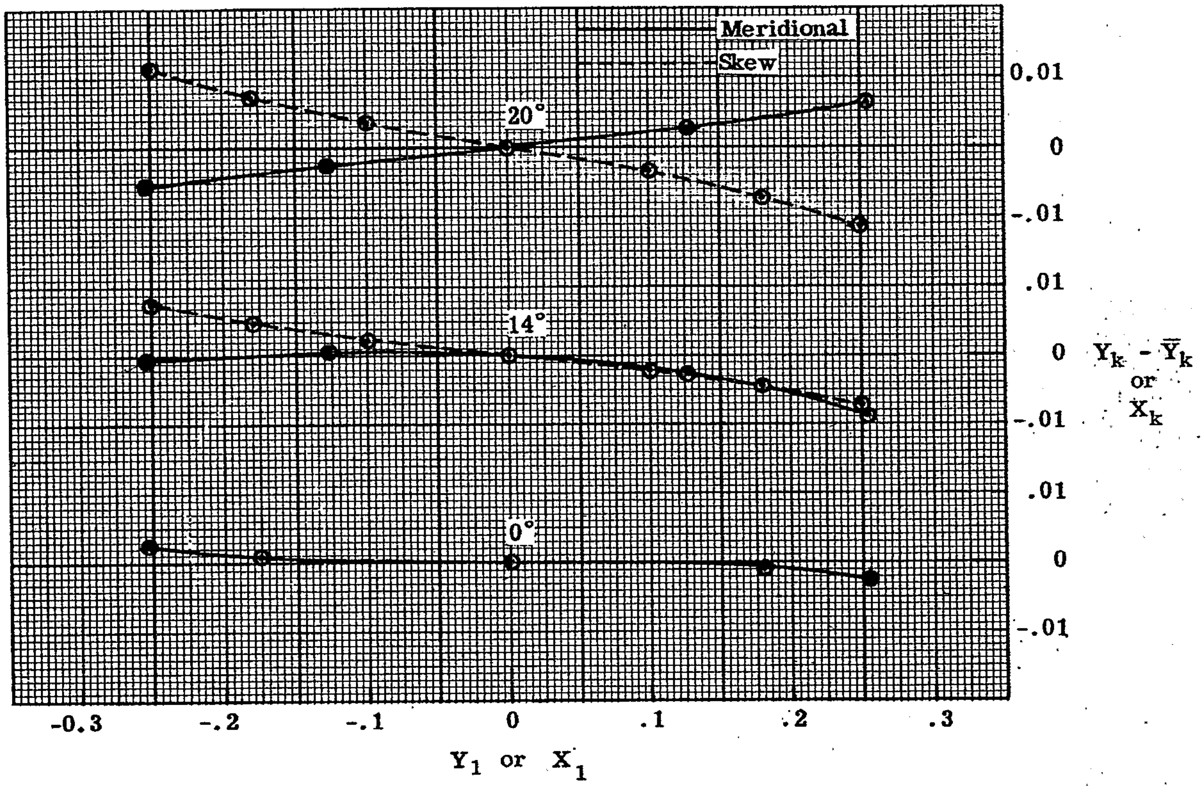


Figure 14.11- Meridional and skew fans for the orthoscopic eyepiece.

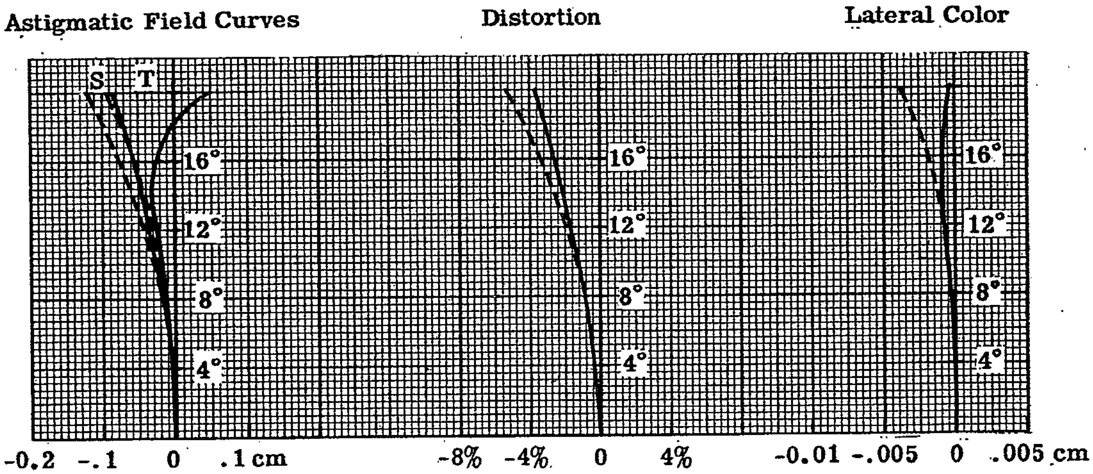
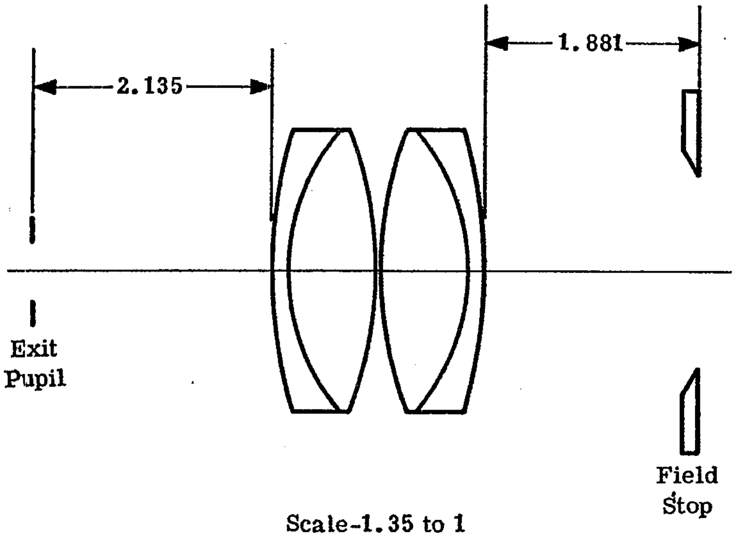


Figure 14.12 - Field, distortion, and lateral color curves for the orthoscopic eyepiece.

14.7 SYMMETRICAL (PLÖSSL) EYEPIECE

14.7.1 Design data. Table 14.5 and Figures 14.13 through 14.15 present the design data and aberration curves for this eyepiece.



Scale-1.35 to 1

Figure 14.13 - Symmetrical (Plössl) eyepiece. Distances in cm.

c	t	Glass Type
0.2135		
	0.1478	649338
0.4868		
	0.8026	517645
-0.2708		
	0.0051	Air (n = 1)
0.2708		
	0.8026	517645
-0.4868		
	0.1478	649338
-0.2135		
	1.881	

$\Sigma P = -0.3013$
$\gamma = 1.307$

Table 14.5 - Lens constants for the Plössl eyepiece. Lengths in cm.

14.7.2 Use and characteristics. This eyepiece, like the orthoscopic, has a long eye relief (about 20 mm) and is well corrected for lateral color. This lens, which has an overall imagery better than that of the orthoscopic, is sometimes used out as far as 25°.

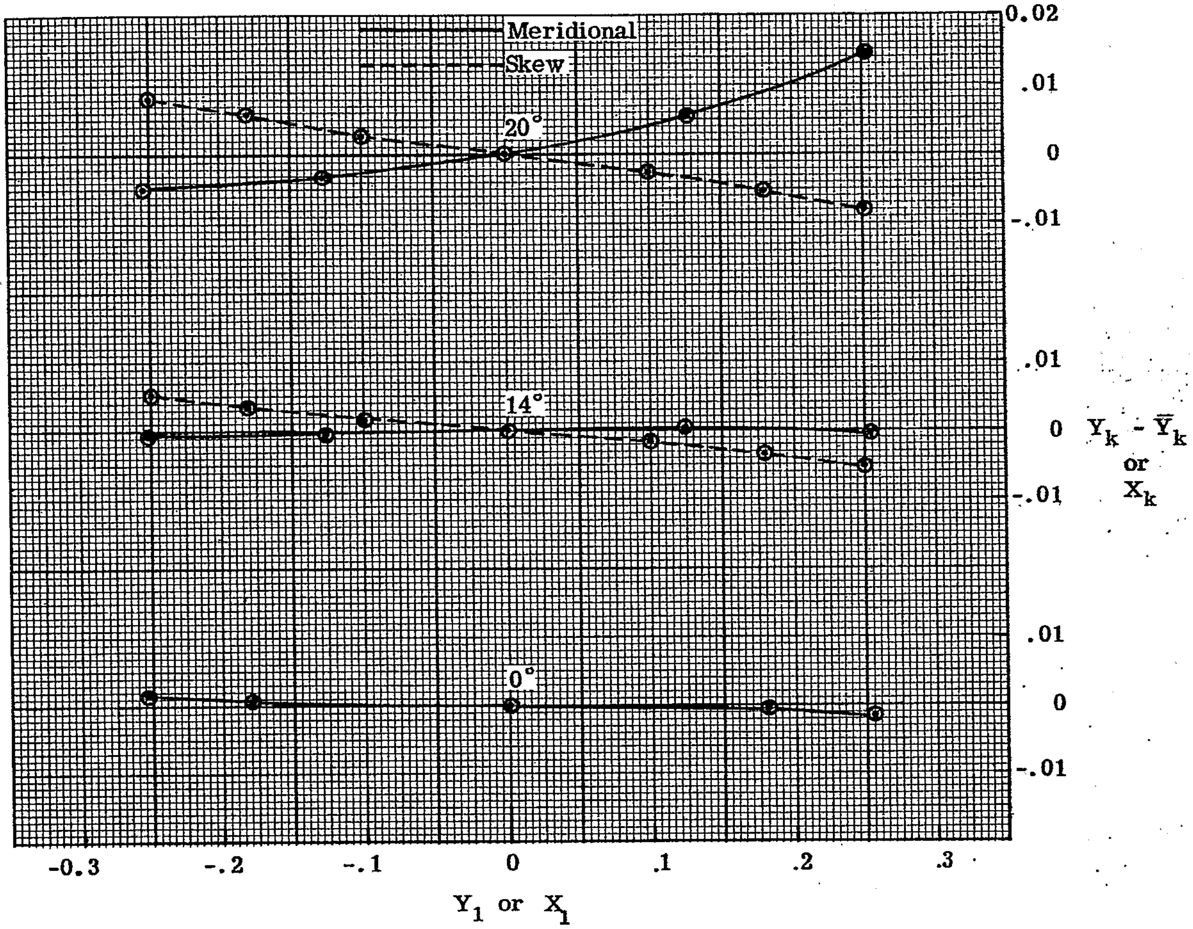


Figure 14.14 - Meridional and skew fans for the symmetrical eyepiece.

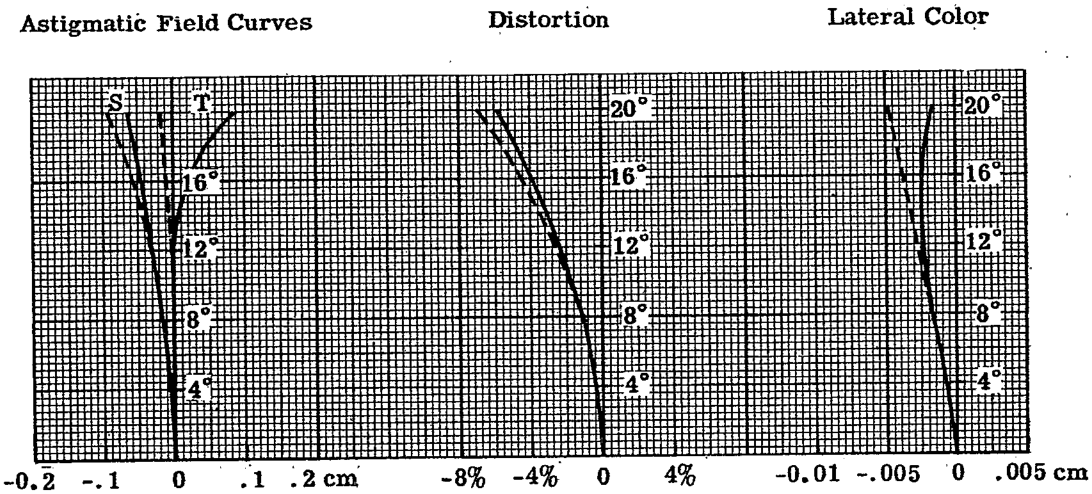
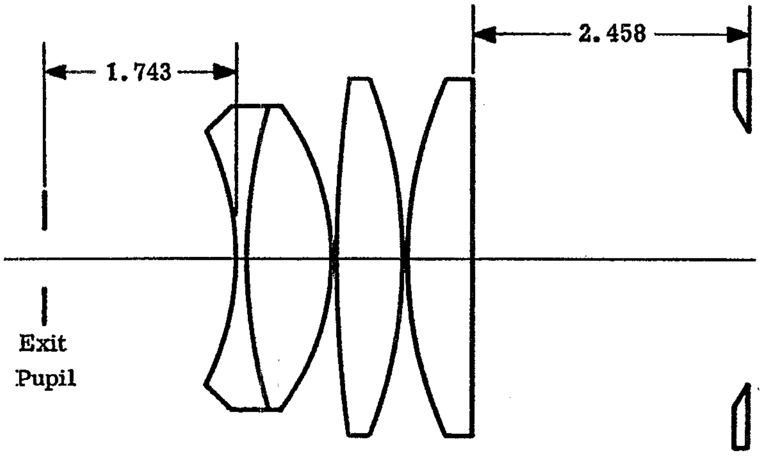


Figure 14.15 - Field, distortion, and lateral color curves for the symmetrical eyepiece.

14.8 THE BERTHELE EYEPIECE

14.8.1 Design data. Table 14.6 and Figures 14.16 through 14.18 present the design data and aberration curves for this eyepiece.



Scale- 1.35 to 1

Figure 14.16 - Berthele eyepiece.
Distances in cm.

c	t	Glass Type
-0.3774		
	0.08	689309
0.2000		
	0.8	620603
-0.4238		
	0.02	Air (n = 1)
0.0714		
	0.60	620603
-0.2107		
	0.02	Air (n = 1)
0.2452		
	0.60	620603
0		
	2.458	

$\Sigma P = -0.2050$
$\gamma = 1.920$

Table 14.6 - Lens constants for the Berthele eyepiece. Lengths in cm.

14.8.2 Use and characteristics. The design aim in this eyepiece is to reduce ΣP, the field curvature. This is accomplished at the expense of lateral color, which is not well corrected.

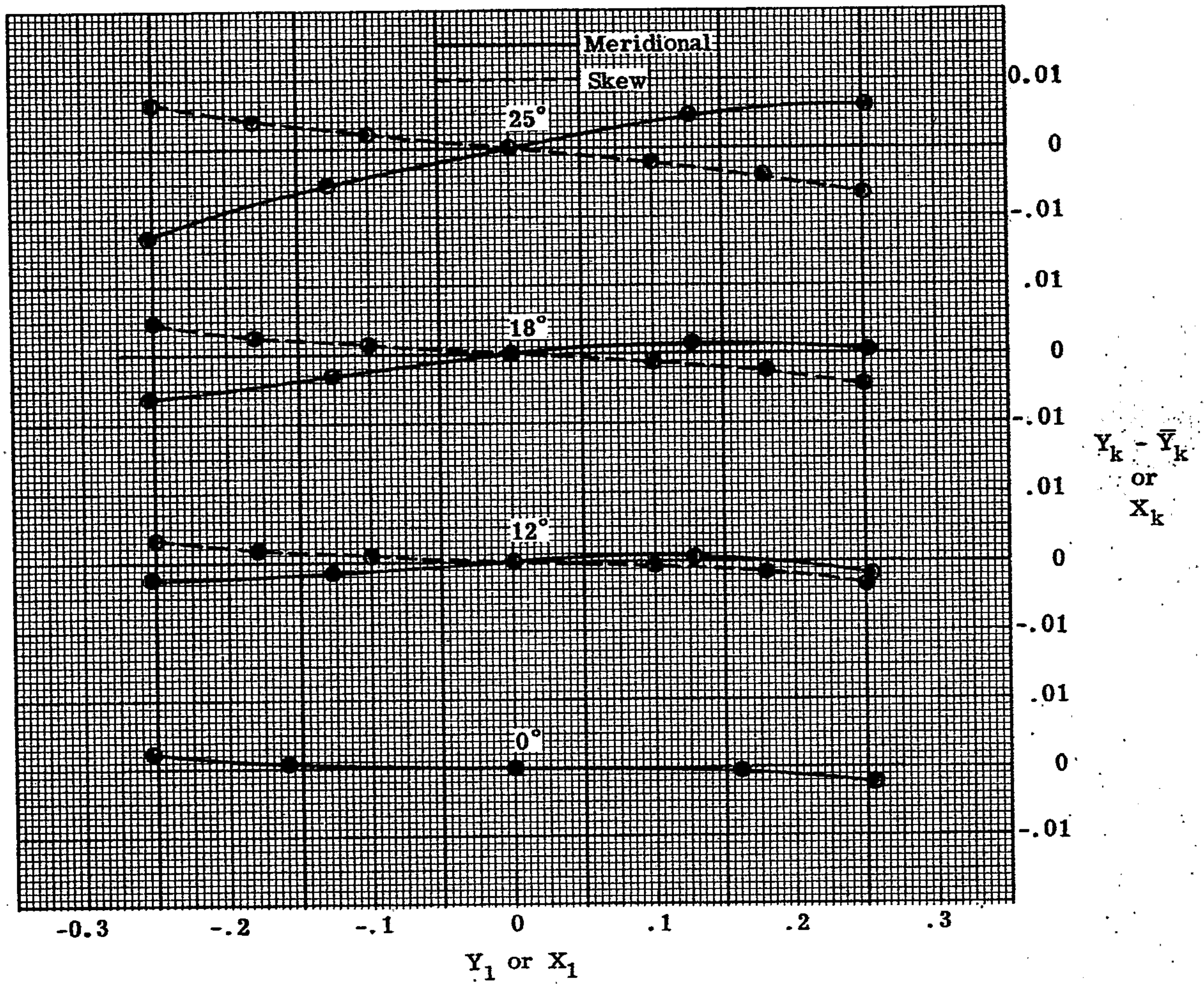


Figure 14.17 - Meridional and skew fans for the Berthele eyepiece.

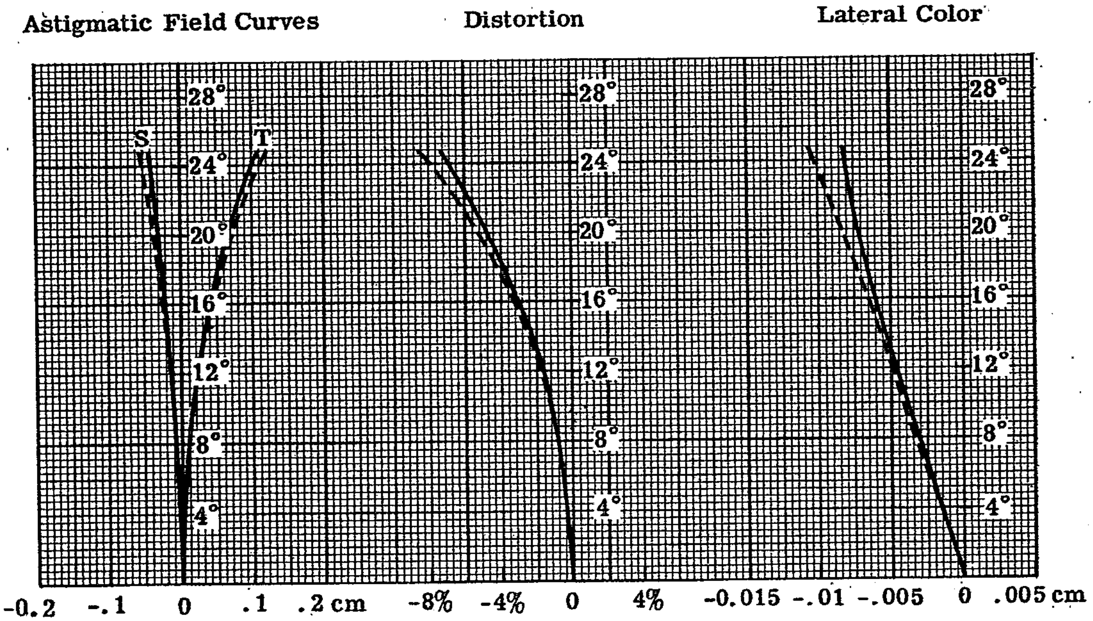
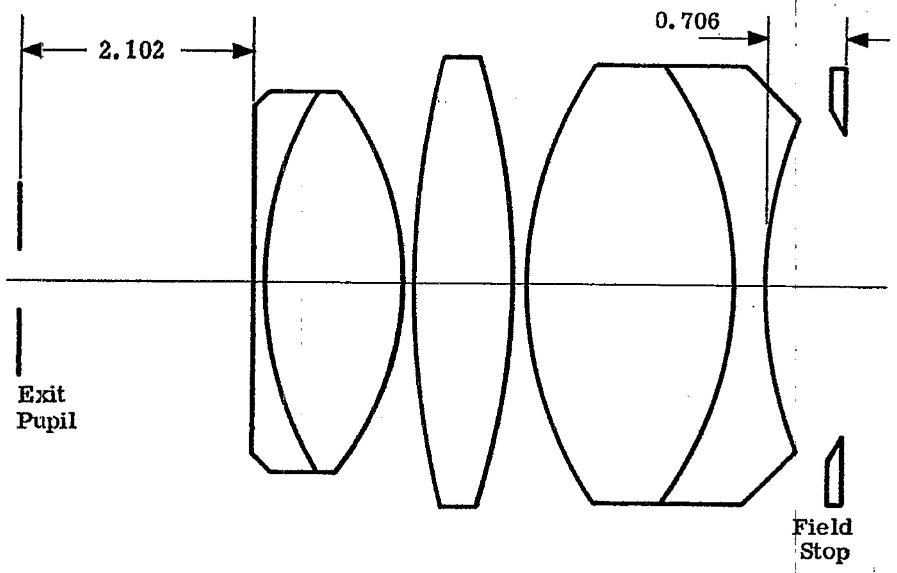


Figure 14.18 - Field, distortion, and lateral color curves for the Berthele eyepiece.

14.9 THE ERFLE EYEPIECE

14.9.1 Design data. Table 14.7 and Figures 14.19 through 14.21 present the design data and aberration curves for this eyepiece.



Scale - 1.35 to 1

Figure 14.19 - Erfle eyepiece. Distances in cm.

c	t	Glass Type
0		
	0.1219	617366
0.2912		
	1.229	517645
-0.3490		
	0.1015	Air(n= 1)
0.1125		
	0.8840	617549
-0.1437		
	0.1015	Air(n=1)
0.2537		
	1.880	611588
-0.2912		
	0.2540	649338
0.2144		
	0.7062	
ΣP = -0.2125 γ = 1.853		

Table 14.7 - Lens constants for the Erfle eyepiece. Lengths in cm.

14.9.2 Use and characteristics. This is a widely used eyepiece which may be designed to cover a half field of 30°. The tangential field curves are controlled fairly well out this far. The lateral color can be corrected better than shown, but one must remember that the eyepiece is used with an objective and prism system. The prisms tend to compensate for the residual lateral color shown here. This is one of the most commonly used wide angle eyepieces. The lateral color is fairly large in the version described, so that sometimes it is designed with an achromatic center lens. The Petzval curvature of the lens is fairly small, but it can be further diminished by reducing the distance between the focal plane and the first surface of the eyepiece. The fallacy with this solution is that any dust on this surface comes sharply into focus. The Petzval curvature can also be reduced by introducing more thickness on the negative lens closest to the exit pupil and by making the surface concave instead of plane. This alternative cuts down on the eye relief.

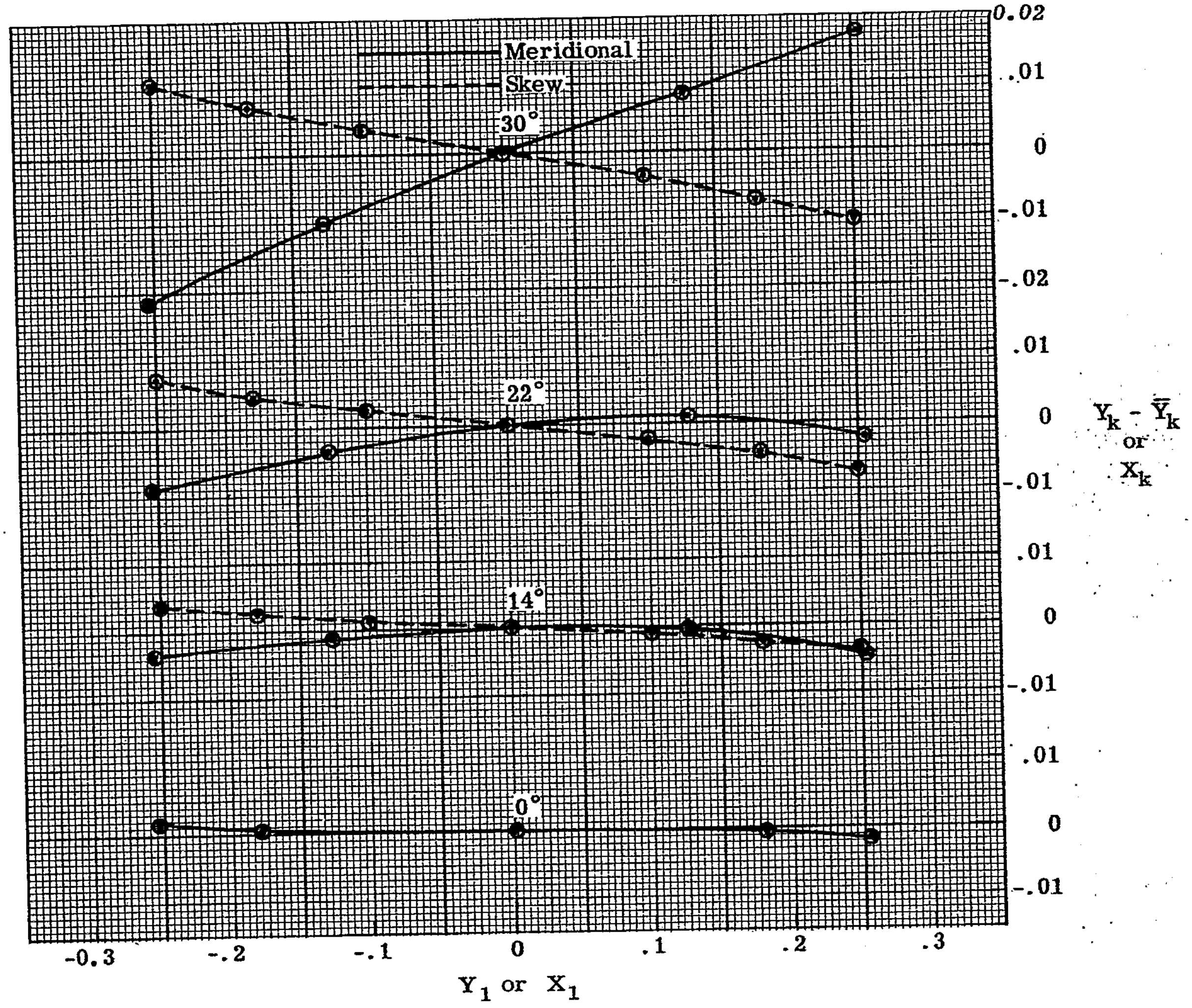


Figure 14.20 - Meridional and skew fans for the Erfle eyepiece.

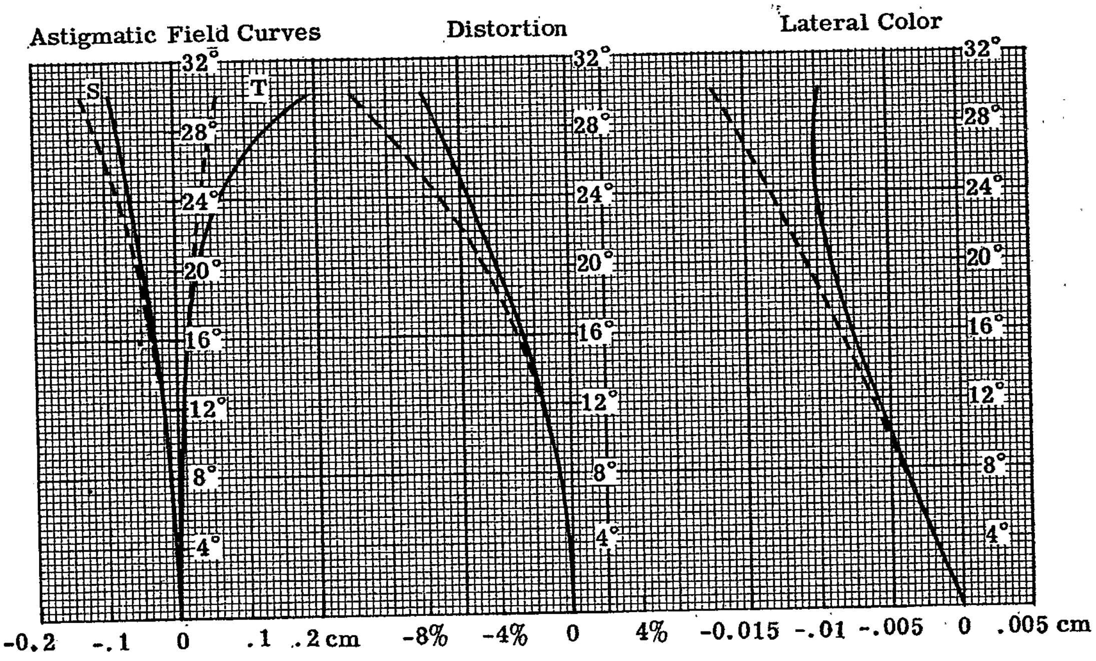
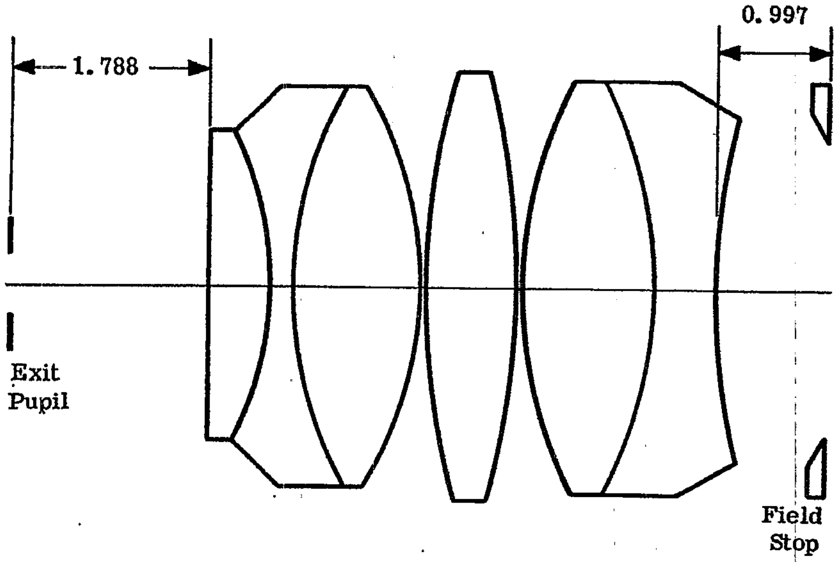


Figure 14.21 - Field, distortion, and lateral color curves for the Erfle eyepiece.

14.10 THE MODIFIED ERFLE EYEPIECE

14.10.1 Design data. Table 14.8 and Figures 14.22 through 14.24 present the design data and aberration curves for this eyepiece.



Scale-1.35 to 1

Figure 14.22 - Modified Erfle eyepiece. Distances in cm.

c	t	Glass Type
0		
	0.55	638555
-0.310		
	0.180	649338
0.275		
	1.15	638556
-0.275		
	0.05	Air(n= 1)
0.123		
	0.80	638555
-0.130		
	0.05	Air(n= 1)
0.234		
	1.20	638555
-0.234		
	0.54	720293
0.159		
	0.9973	

$\Sigma P = -0.221$
$\gamma = 1.78$

Table 14.8 - Lens constants for the modified Erfle eyepiece. Lengths in cm.

14.10.2 Use and characteristics. This eyepiece is an improvement on the Erfle eyepiece. The lateral color is better and the tangential and sagittal fields are not as widely split. It still has a good eye relief. The distortion is large but for telescopes this is not too objectionable because the field stop is round. Hence the corners of the field, which suffer from large distortion, are missing.

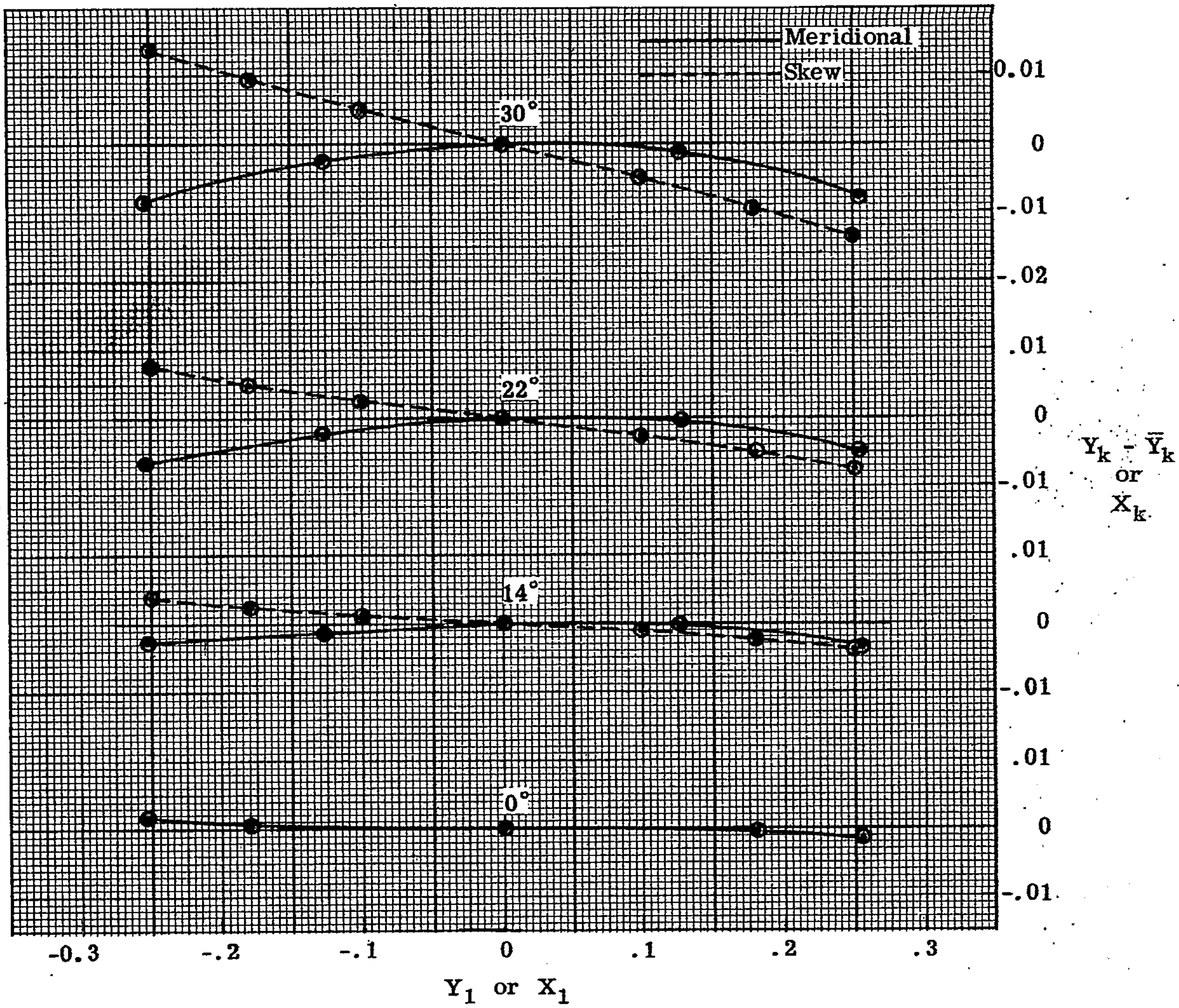


Figure 14.23 - Meridional and skew fans for the modified Erfle eyepiece.

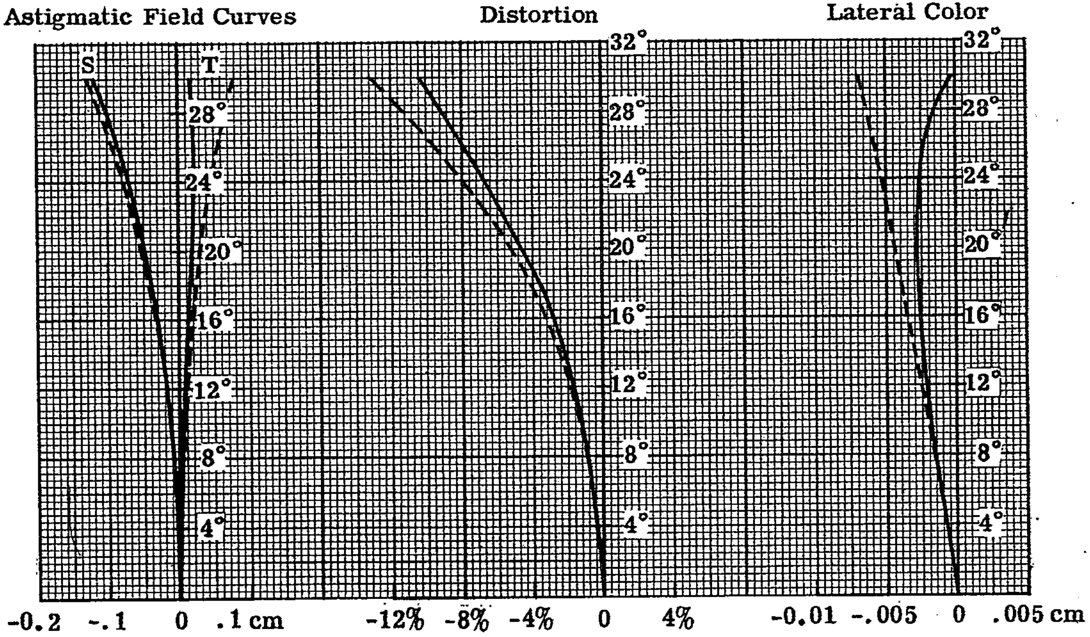
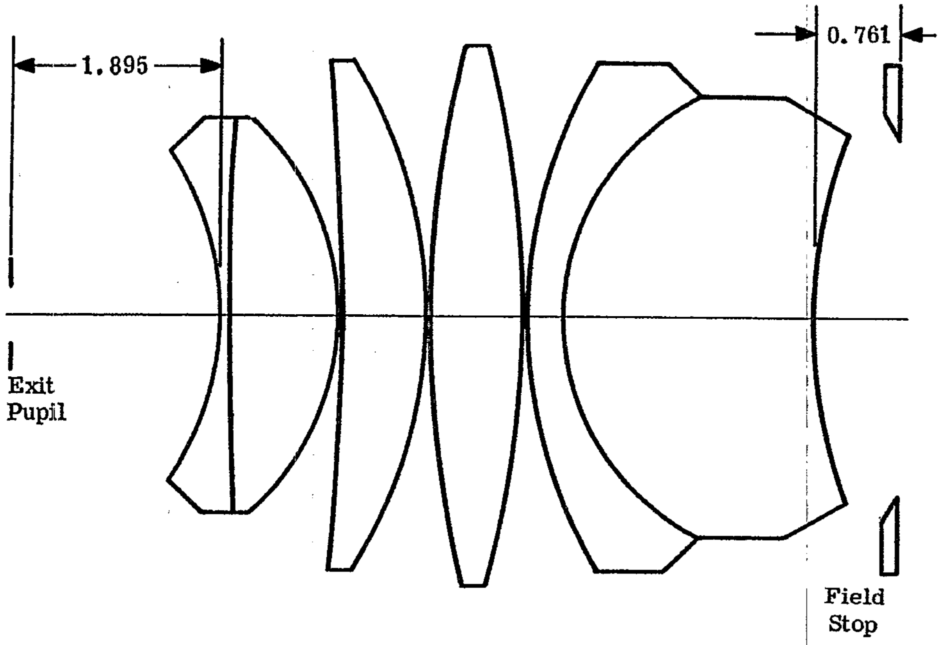


Figure 14.24 - Field, distortion, and lateral color curves for the modified Erfle eyepiece.

14.11 THE WILD EYEPIECE

14.11.1 Design data. Table 14.9 and Figures 14.25 through 14.27 present the design data and aberration curves for this eyepiece.



Scale - 1.35 to 1

Figure 14.25- Wild eyepiece. Distances in cm.

c	t	Glass Type
-0.3636		
	0.10	689309
0.04		
	0.95	620603
-0.4000		
	0.01	Air(n=1)
-0.0200		
	0.80	620603
-0.2243		
	0.01	Air(n=1)
0.1025		
	0.85	620603
-0.1025		
	0.01	Air(n=1)
0.2171		
	0.35	649338
0.4400		
	2.25	573425
0.2170		
	0.761	
ΣP = -0.1538 γ = 2.56		

Table 14.9-Lens constants for the Wild eyepiece. Lengths in cm.

14.11.2 Use and characteristics. This rather complex eyepiece is interesting because the Petzval curvature is so small. The tangential field is also well under control out as far as 36°. The Petzval curvature is kept small by using strongly curved surfaces as the outside surfaces of the lens. If this is done the glass used for the element nearest the field stop must be free of bubbles; otherwise they will be seen since they are so close to the focal plane.

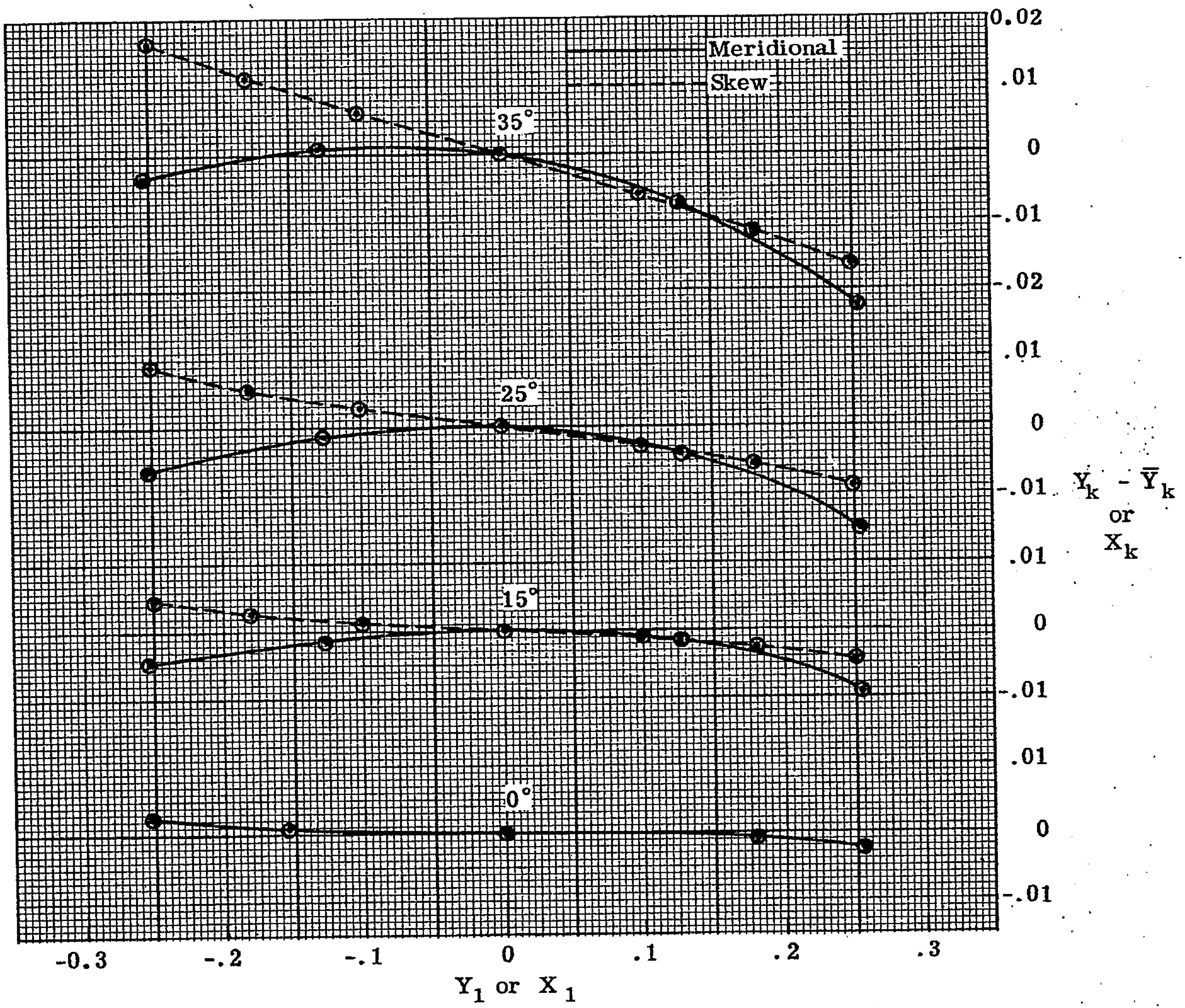


Figure 14.26- Meridional and skew fans for the Wild eyepiece.

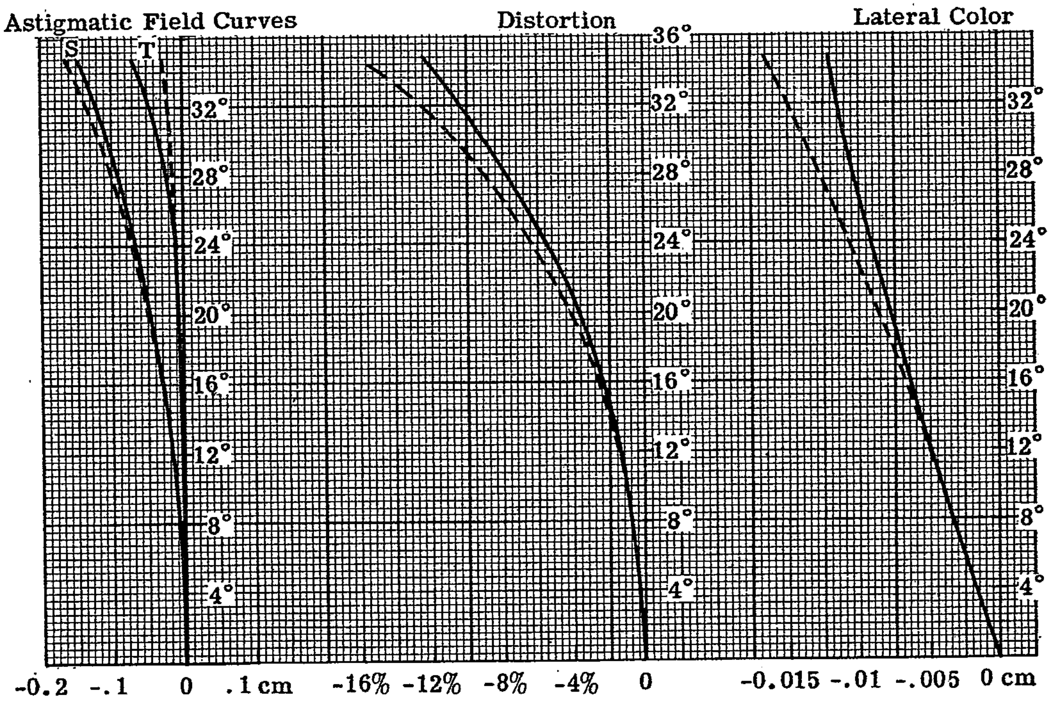


Figure 14.27- Field, distortion, and lateral color curves for the Wild eyepiece.

14.12 SUMMARY

The eyepieces shown in the previous sections are merely representative of types. When they are used in differing applications slight modifications should be made to correct the aberrations of the system. In Section 15 a telescope with prism system is designed to show how the eyepiece is adjusted to fit the particular problem.

15 COMPLETE TELESCOPE

15.1 INTRODUCTION

The problems encountered in the design of a complete telescope are well suited to illustrate how the individual design of objective, erecting system, and eyepiece must be fitted into the overall solution. If the limits on space, vibration, method of support, and other factors permit or demand it, a lens type erecting system, rather than a prism system, may be employed. The basic concepts of this type were discussed in Section 7.5.3. The refinement of the lens design is patterned after the techniques described for the objective and eyepiece. However, the designer is usually faced with restrictions on space and other considerations which require that he fold the light path. Let us therefore consider such a case.

15.2 THE DESIGN PROBLEM

Suppose the following specifications are established for a telescope:

(1)	Magnifying Power	10 X
(2)	Apparent Field of View	30° (half angle)
(3)	Exit Pupil Diameter	0.5 cm.
(4)	Minimum Eye Relief	2.0 cm.
(5)	Line of sight to be displaced a minimum of one inch in a plane at 45 degrees to the observer's vertical and to his right. (This is actually a conclusion drawn from more complex requirements but will serve to establish the need for a displaced line of sight.)	

15.3 PRELIMINARY CONSIDERATIONS

15.3.1 Prism type. From requirement (5) above and from Section 13.10.2 we can easily see that a Porro prism system will offer a ready solution to displacement and erection if we have A $\geqq$ 0.7 inch (approx.)

15.3.2 The eyepiece.

15.3.2.1 From Section 14 we can also quickly determine that it will be necessary to use an eyepiece of the Erfle type, since, from requirement (2), the apparent field must be 30°.

15.3.2.2 We now must determine the focal length of the eyepiece. One can say almost without qualification that the longer the focal length of the eyepiece, the better the image quality of the system. Usually, however, this means the telescope will become large, expensive, and cumbersome. Most commercial applications call for a small compact system. There is, however, a lower limit to the focal length of the eyepiece, since there is a minimum eye relief which can be used with comfort by the observer. The data on the Erfle eyepiece (Section 14.9) showed that the eye relief is around 0.8 of the focal length. Therefore in order to meet requirement (4) it will be necessary to have an eyepiece focal length of at least 2.5 cm.

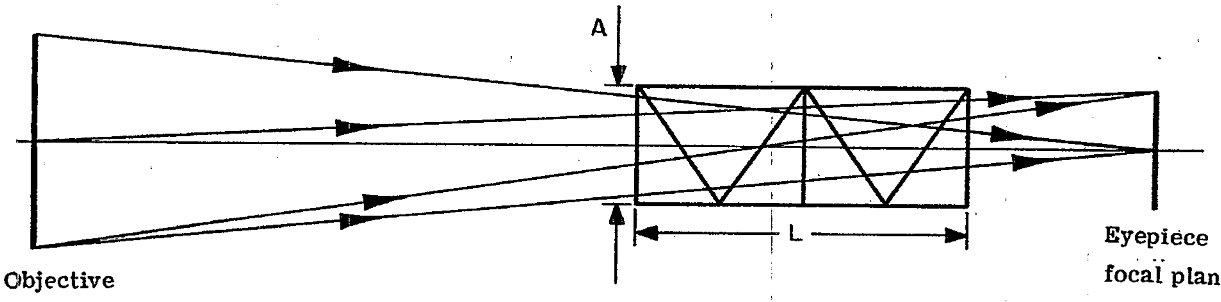
15.3.3 Preliminary summary. We have now established the following design parameters.

(1) Objective:	(a)	Focal length, f_o	25 cm.
	(b)	Diameter	5.0 cm.
	(c)	f-number	5.0
	(d)	Field of view (half angle)	3° (approx.)
(2) Eyepiece:	(a)	Type	Erfle
	(b)	Focal length, f_e	2.5 cm. (min)
	(c)	Eye relief	2.0 cm. (min)
	(d)	Field of view (half angle)	30°

(3) Erecting System:	(a)	Type	Porro Prism System
	(b)	Aperture	1.8 cm. (min.)

15.4 DESIGN REFINEMENT

15.4.1 The Porro erecting system. The next step in the design is to determine the size of prisms needed to erect the image. The prism length is determined by drawing a thin lens telescope system and using a tunnel diagram for the prisms. In order to keep the prism small, as much as 50% vignetting is allowed at the edge of the field. The glass type selected for the prisms is of importance. It must have a high transmission with relatively low dispersion, and the index of refraction must be high enough to insure total reflection for all the rays. A glass frequently used in prisms is type number 573574. The drawing shown in Figure 15.1 shows the layout of the prism system used in this sample problem. The prism aperture is 2.9 cm and the total thickness of the prism is 11.6 cm.



L' = Reduced prism length = 4 A/n = 7.3767 cm.
A = Aperture diameter of prisms.
n = Index of refraction of prisms.

Figure 15.1 - Diagram illustrating positioning of Porro prism system in telescope. The prisms are shown "reduced".

15.4.2 The objective.

15.4.2.1 The objective design is started by consulting Table 11.3 for a thin lens solution. In this example the following solution (case No. 14) was used:

Lens	(a)	517645
	(b)	689309

15.4.2.2 From the curvatures given in the table, it is possible to draw up the lens and assign the proper thicknesses. Figure 15.2 shows a scale drawing of the thin lens solution (curvatures) with proper thickness added. Then, with the thicknesses and the prism added, the third order aberrations are computed to compensate for the eyepiece.

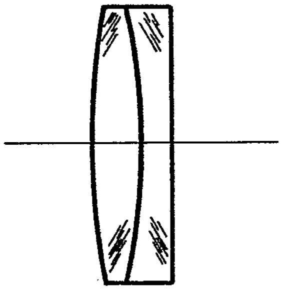


Figure 15.2 - Scale drawing of the objective.

15.4.3 The eyepiece

15.4.3.1 The preliminary eyepiece may be scaled from the Erfle shown in Section 14.9. The astigmatism and lateral color will probably have to be adjusted to match the objective and prisms. They may be controlled either by changing the curve on the cemented surfaces or by changing the glass in the eyepiece. The astigmatism and lateral color are corrected in the eyepiece, and then the total coma and spherical aberrations are corrected in the objective. Very little can be done about the distortion introduced by the eyepiece. It is closely connected to the astigmatism, so once the astigmatism is corrected to the desired value, the distortion is determined. Distortion is really not a very objectionable aberration in a viewing telescope.

15.4.3.2 The lateral color is made more positive by finding glasses with reduced dispersion for the positive elements. If the index difference between the positive and negative lenses of the cemented doublets is reduced, it is possible to make the cemented curves shorter in radius (thereby adding positive lateral color) without introducing high order positive astigmatism.

15.4.3.3 In correcting the astigmatism in the eyepiece, it is necessary to ray trace for every change. The reason for this is that near the edge of the field the astigmatism is dominated by higher than third order aberrations.

15.4.4 Objective readjustment. After the astigmatism and the lateral color have been corrected to match those of the objective and the prisms, it is necessary to readjust the objective to correct for the axial color, spherical aberration and coma of the complete telescope.

15.5 THE COMPLETED DESIGN.

The completed telescope system, shown in Table 15.1, represents a solution to the design problem. The aberration curves are shown in Figures 15.4, 15.5, and 15.6. Figure 15.4 is a plot of the angular aberrations in D light for skew fans of rays at three obliquities. Figure 15.5 is a similar plot for meridional fans of rays. Figure 15.6 contains field curves, a plot of distortion, and lateral color curves. In all three figures, the dashed curve represents the third order. This telescope was corrected by an expert designer. It represents excellent correction, so it may be used as a guide on what to expect from such a telescope. Note in Figure 15.6 how the final T and S curves are adjusted. At the edge of the field they are split by 3.7 diopters. The mid-focus is inside the paraxial focus by 0.8 diopter. This means if the eyepiece is focused in by 0.8 diopters, the image quality will essentially be free of astigmatism out to 20° . These aberrations may appear to be very large but they are typical and are not as objectional as it may seem. A telescope is used for acute vision primarily close to the axis (within ±12° apparent field). The observer seldom uses the telescope in a fixed position and rolls his line of sight around to observe objects near the edge of the field. The edge of the field is usually used to notice motion. If anything of interest does appear in the edge of the field, the observer can train the telescope to center it in the field. When an observer has his eyes to the telescope he wants all the field of view he can have. It is much better to have a picture blurred at the edges than none at all. For this reason a telescope should always be designed for as wide a field as possible, even if the astigmatism and distortion may become large. The limit should be set by the size of the instrument and the cost, rather than by the image quality. As the field is increased beyond the 30° half angle of the Erfle, the size of the instrument grows rapidly, for the prisms and eyepiece must be enlarged. In wide angle telescopes it is also desirable to maintain as large an exit pupil as possible. The reason for this is that the iris of the observer's eye is not located at the center of rotation of the eye. With the iris located at the exit pupil there is a tendency for the iris to rotate out of the exit pupil when viewing objects near the edge of the field. This is demonstrated in Figure 15.3. It is true that the observer may move his eye but in a binocular instrument this is not possible for both eyes.

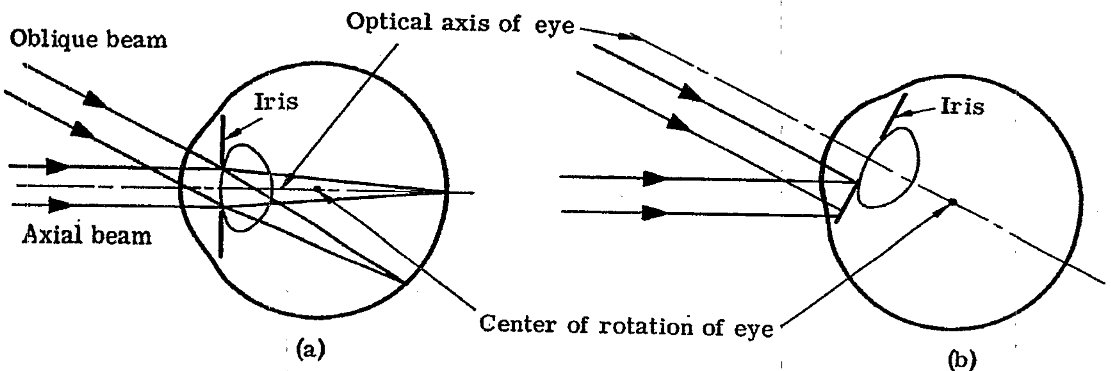


Figure 15.3 - Diagram (a) illustrating the eye viewing an axial image. Diagram (b) illustrating the eye rotated to view an oblique image and losing the entire beam.

c	t	n	ν
0.07080			
	0.80	1.517	64.5
-0.06874			
	0.05	1.00	
-0.07165			
	0.40	1.689	30.9
-0.02413			
	15.00	1.00	
0.0000			
	11.60	1.5725	57.4
0.0000			
	3.0186	1.00	
-0.21670			
	0.254	1.649	33.8
0.35000			
	1.8796	1.620	55.5
-0.25000			
	0.1015	1.00	
0.15350			
	0.8840	1.620	55.5
-0.10000			
	0.1015	1.00	
0.35000			
	1.2294	1.517	64.5
-0.27500			
	0.12190	1.617	36.6
0.00000			

Table 15.1 - Specification for 10 X telescope.

15.5.1 Eyepiece and objective checks. In the preliminary design it is advisable to correct the eyepiece and objective as separate units. Usually designers trace parallel rays into the eyepiece from the eyeside towards the focal plane, and trace parallel rays through the objective to the focal plane. The transverse image errors are then made to match at the intermediate focal plane. After it appears that the two match reasonably well they should be put together and studied as a complete telescope. It is advisable to insert a dummy reference surface at the internal focal plane so that when the rays are traced through the entire system it will be possible to note the image errors on the image plane.

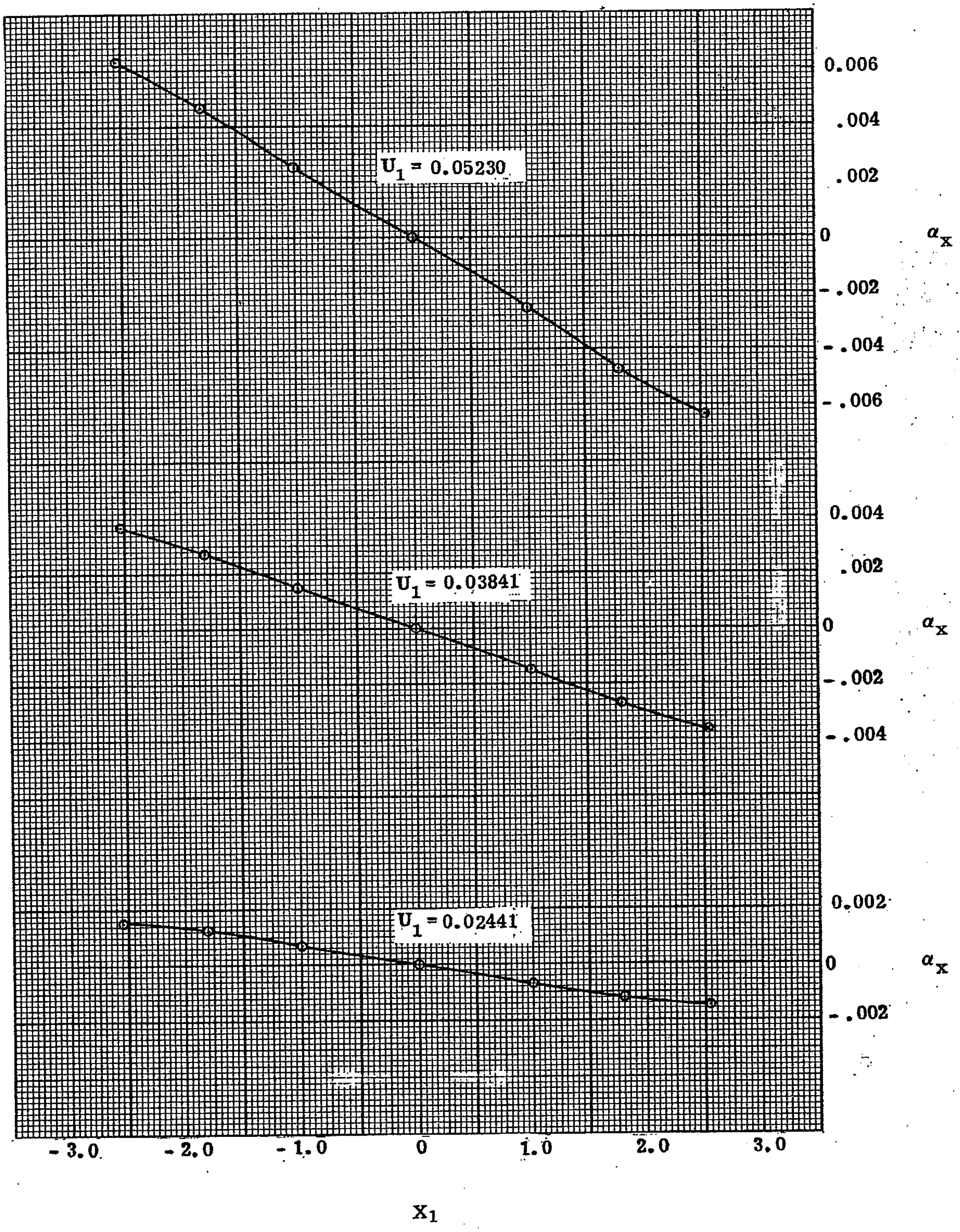


Figure 15.4 - Skew rays.

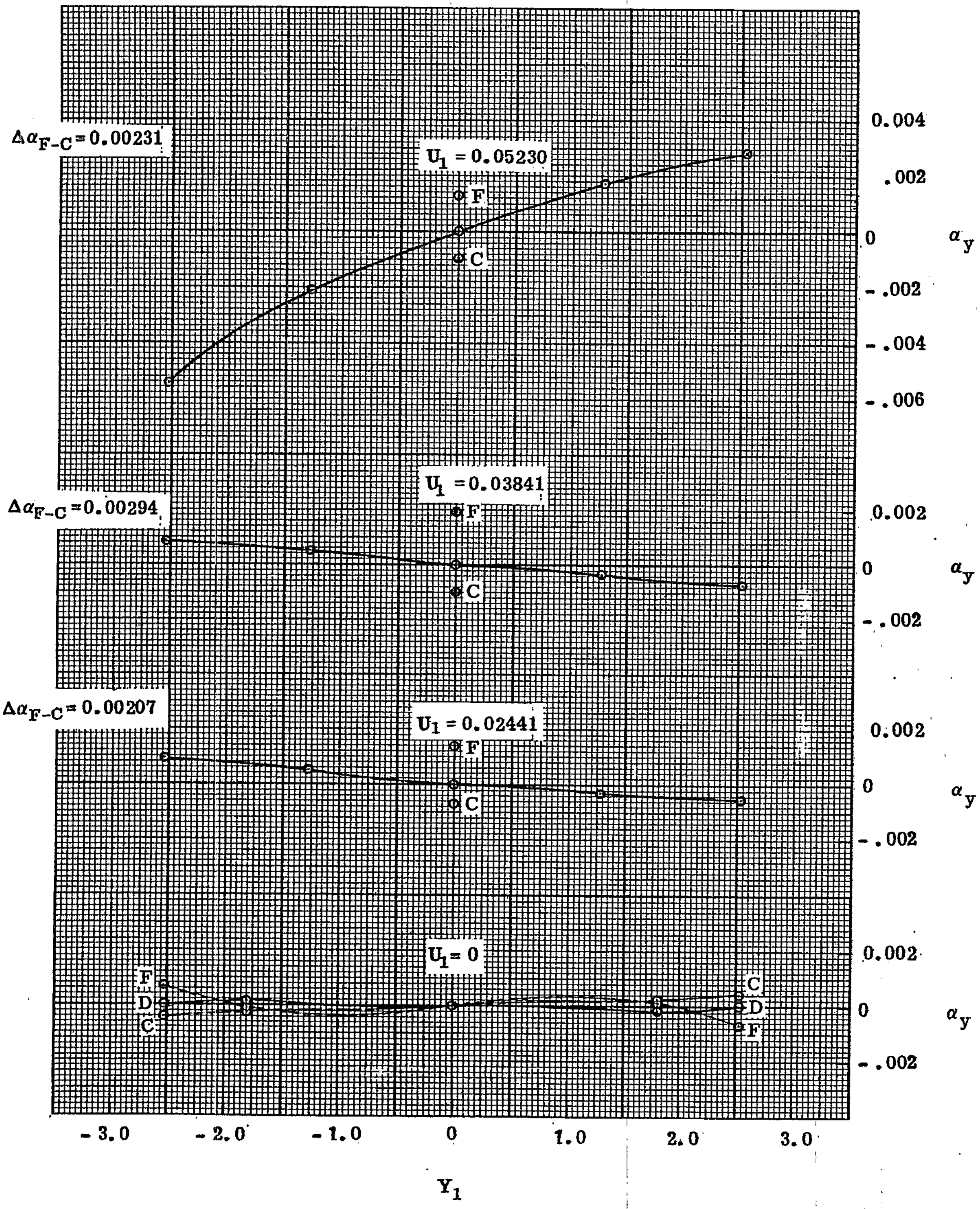


Figure 15.5 - Tangential rays.

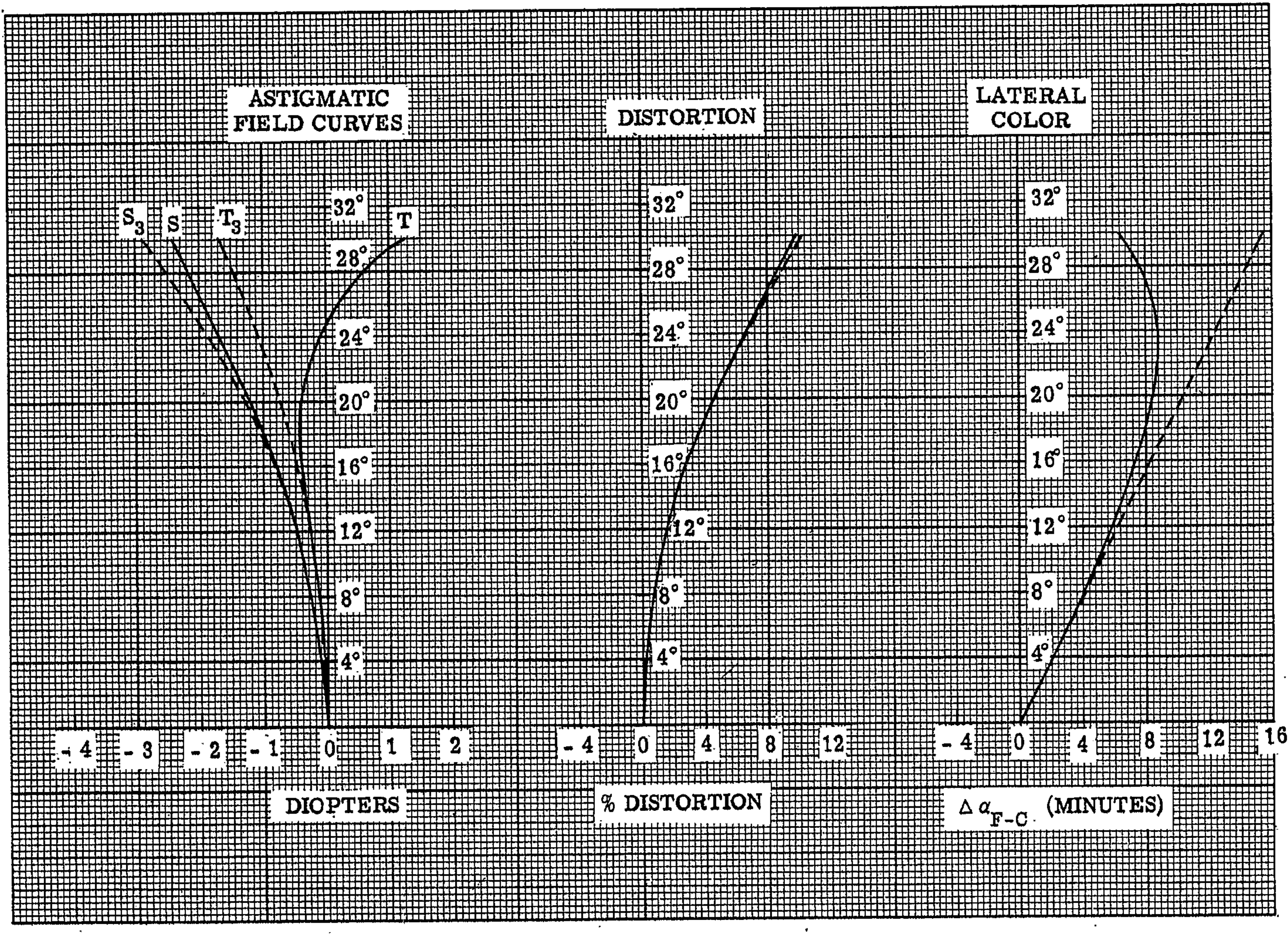


Figure 15.6 - Astigmatic field curves, distortion, and lateral color.

16 APPLICATIONS OF PHYSICAL OPTICS

16.1 INTRODUCTION

16.1.1.1 Restatement of principles. In instruments for purposes of interferometry, the problems of geometrical optical designs are usually simple. However, since such instruments depend upon the interference of light waves for their proper functioning, a knowledge of the principles of interference is necessary for proper design. These principles have been already presented. A brief recapitulation of these principles is now presented, followed by detailed examples of their application to the design of several typical instruments of this class.

16.1.1.2 As stated in Section 3, the instantaneous magnitude of a plane-polarized light wave will be equivalent to the instantaneous magnitude of the electric vector and can be specified by the trigonometric function

$$E(z, t) = a \cos(knz + \phi - \omega t) \qquad (1)$$

where

z = distance measured along Z
t = time
k = $2\pi/\lambda$
ω = $2\pi/T$
λ = wavelength
T = period for one complete vibration
ϕ = phase angle
n = refractive index. It can be a function of z for variable media.
a = amplitude of the wave. It is an exponential decreasing function of z for absorbing media.

It is also shown that the time-averaged energy density for a single wave over a single period T of oscillation will be proportional to the square of the amplitude, that is,

$$W = a^2/2. \qquad (2)$$

16.1.1.3 If interference phenomena of two or more waves are considered, the time-averaged energy density, W, will be the sum of the instantaneous energies of the electric vectors, the average over a single period of T of the square of the sum of the instantaneous magnitudes of the electric vectors. Thus, for two collinear waves,

$$W = \frac{1}{2}\left[a_1^2 + 2a_1 a_2 \cos(\phi_1 - \phi_2) + a_2^2\right] \qquad (3)$$

where

ϕ_1, ϕ_2 = phase angles of each electric vector
$\phi_1 - \phi_2$ = fixed phase difference, δ.

16.1.1.4 Referring again to Section 3, the conditions of Equation (3) depend on the direction of propagation and the source of radiation. Collinear, coherent waves will reinforce each other when the phase difference is zero or an even multiple of π and oppose each other when the phase difference is an odd multiple of π. For collinear, non-coherent waves, this reinforcement or opposition does not apply, but the time-averaged energy densities will add according to

$$W = \frac{1}{2}\left[a_1^2 + a_2^2\right]. \qquad (4)$$

16.1.1.5 If the waves are non-collinear and coherent, as if Figure 16.1, their phases Φ_1 and Φ_2 will be given by

$$\begin{aligned} \Phi_1 &= knz + \phi_1 \\ \Phi_2 &= kn(x \sin\theta + z\cos\theta) + \phi_2 \end{aligned} \qquad (5)$$

where

ϕ_1 = phase angle of the wave propagated along OZ,
ϕ_2 = phase angle of the wave propagated along OP.

The difference in phase angles will then be

$$\Phi_1 - \Phi_2 = \varphi_1 - \phi_2 - knx\sin\theta + knz(1 - \cos\theta). \qquad (6)$$

Letting $\phi_1 - \phi_2 = \delta$ and using Equation (3), the time-averaged energy density will be

$$2W = a_1^2 + a_2^2 + 2\,a_1\,a_2 \cos\left[\delta - knx\sin\theta + knz\,(1 - \cos\theta)\right] \qquad (7)$$

where a_1 and a_2 are the amplitudes of the interfering waves at the point $(0, y, 0)$. By choosing θ to be suitably small, we can set $\sin\theta = \theta$ and $1 - \cos\theta = \theta^2/2$. If observation is to be made in the xy plane near $z = 0$, the z term can be neglected and Equation (7) becomes

$$2W = a_1^2 + a_2^2 + 2\,a_1\,a_2 \cos\left(\delta - \frac{2\pi nx\theta}{\lambda}\right) \qquad (8)$$

which is the usual interference formula.

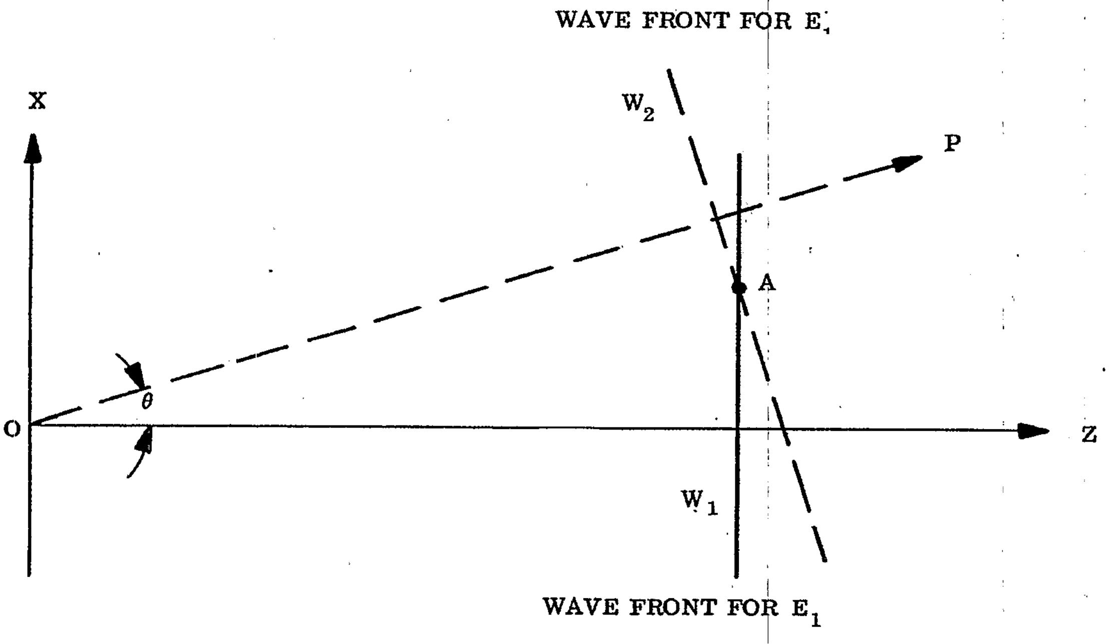


Figure 16.1 - Interference between two plane wavefronts W_1 and W_2 that are propagated along different directions.

16.2 THE FIZEAU INTERFEROSCOPE

16.2.1 Principles of operation.

16.2.1.1 A group of interferometers known as Fizeau interferoscopes or Fizeau double beam interferometers have been devised around afore mentioned principles for the purpose of testing the flatness and parallelism of the surfaces S_1 and S_2, Figure 16.2, of a plane parallel plate or for testing the flatness of a surface against an optical flat. The essential characteristics of these interferometers are illustrated in Figure 16.2 in which either of the surfaces S_1 or S_2 may be the optically flat surface of reference. Monochromatic light issues from a pinhole H and emerges from the collimator L_1 as plane waves. By a slight angular adjustment (not shown) of the upper plate surface, S_2 can be made to reflect ray HA approximately back upon itself. We take this direction as the OZ-direction. Surface S_1 now reflects ray HAB along a direction BR such that angle $\theta = 2\alpha$ where α is the indicated angle between surfaces S_1 and S_2. We choose the direction OP parallel to BR. The coordinate line OX falls in the wave front reflected by S_2. Lines OP, OX and OZ together with the angle are now corresponding elements in Figures 16.1 and 16.2. Equation (7) or (8) may therefore be applied to determine the energy densities at any point (x, z). If both surfaces S_1 and S_2 are flat, σ is constant; but we must take $\sigma = \sigma(x)$ when both surfaces are not flat. In the interests of simplicity, we shall suppose at first that surfaces S_1 and S_2 are flat.

16.2.1.2 When S_1 and S_2 are uncoated surfaces of glass, the two waves formed by reflection at S_1 and S_2 will have amplitudes a_1 and a_2 so nearly alike that one may set $a_1 = a_2 = a$ and write Equation (7) as

$$W = a^2 \left\{ 1 + \cos \left[\delta - knx \sin\theta + knz (1 - \cos\theta) \right] \right\}. \qquad (9)$$

Furthermore, $\delta = \pi$ for optically flat surfaces of glass since the phase changes due to reflection at A and B differ by 180°. Thus,

$$W = a^2 \left\{ 1 - \cos kn \left[x \sin\theta - z (1 - \cos\theta) \right] \right\} \qquad (10)$$

if the surfaces S_1 and S_2 are optically flat surfaces of glass.

16.2.1.3 Lens L_1 and L_2 are invariably arranged so that the plane $z = 0$ or a neighboring plane is focused upon the retina or upon the photographic emulsion, i.e., one arranges to observe the energy density W in the interference fringes in a plane for which z is either zero or small. Also, the angle θ is very small in actual practice. Thus both z and $1 - \cos\theta$ become so small that one will ordinarily be justified in neglecting the term $z(1 - \cos\theta)$ in Equation (10), and in writing

$$W = a^2 \left[1 - \cos kx\theta \right] = 2a^2 \sin^2 \left(\frac{\pi x \theta}{\lambda} \right)$$

$$= 2a^2 \sin^2 \left(\frac{2\pi x \alpha}{\lambda} \right) \qquad (11)$$

when the space between S_1 and S_2 is air.

16.2.1.4 The actual energy density W is of little interest in practical interferometry. Interest centers, rather, upon the fringe-width, the distance from one fringe to the next similar interference fringe. The fringe system is repeated, according to Equation (11), whenever x is altered by the amount Δx such that $2\pi\alpha \, \Delta x / \lambda = \pi$, i.e., whenever

$$h = |\Delta x| = \frac{\lambda}{2\alpha} \qquad (12)$$

where h denotes fringe-width and α is the angle in radians between surfaces S_1 and S_2. Equation (12) can be used to measure α. If $\alpha = 0$, the fringe-width is infinite, and conversely.

16.2.1.5 It will be seen from Figure 16.2 that

$$d = x\alpha \qquad (13)$$

where d is the thickness of the air gap at point x. Equation (11) can therefore be written in the highly instructive form

$$W = 2a^2 \sin^2 \left(\frac{2\pi d}{\lambda} \right) \qquad (14)$$

Hence W is constant for those loci along which the separation d of the surfaces is constant. W is, of course, constant along an interference fringe. Therefore, each interference fringe is the locus of points for which the separation d of the surfaces is constant. This statement holds throughout interferometry with very few exceptions or qualifications. With respect to Equation (14), we note that W has the period $d = \lambda/2$. This means that in the Fizeau interferoscope the separation d changes by $\lambda/2$ in going, say, from one bright fringe to the

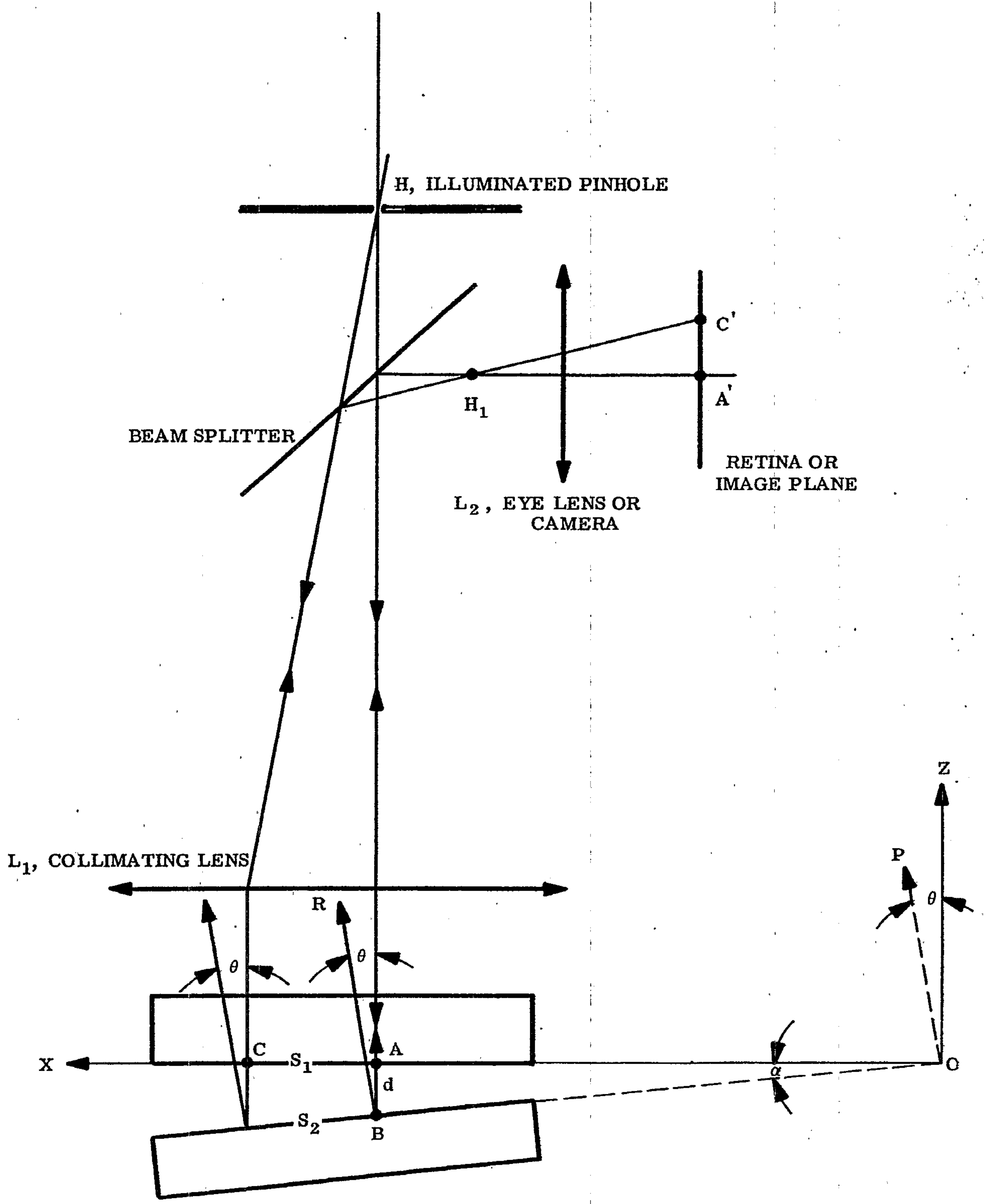


FIGURE 16.2 -Notation with respect to the Fizeau Interferoscope or Interferometer.

next. More generally, it is the optical path nd that changes by $\lambda/2$.

16.2.1.6 The use of the Fizeau interferoscope for examining parallelism of plates amounts to considering the gap between surfaces S_1 and S_2, Figure 16.2, as the plate. The refractive index of the gap is now that of the plate.

16.2.2 The Fizeau interferoscope in testing for optical flatness.

16.2.2.1 The separation d of surfaces S_1 and S_2, Figure 16.2 can be large. Consequently the risk of scratching a surface during testing for flatness is avoided. Moreover, the reference flat is not subjected to wear by frequent rubbing, etc.

16.2.2.2 As can be expected, the interference fringes will not be straight unless the test surface is also an optical flat.

16.2.2.3 In paragraph 16.2.1.2 we saw that $\delta = \pi$ for optically flat, uncoated surfaces S_1 and S_2 of glass. The main effect of a small departure of the test surface from a plane is to introduce a local irregularity in the separation d, Figure 16.2, over the range of x at which the departure occurs. It is natural, then, to consider δ in the form

$$\delta = \pi - 2kn\,D(x) \tag{15}$$

where D(x) shall express the phase difference introduced between the two interfering waves on account of the departure of the reflecting surfaces from a plane. We suppose, as in the argument leading to Equation (11), that the term $knz\,(1 - \cos\theta)$ is negligible in Equation (9) and introduce δ from Equation (15). The result is

$$W = a^2 \left[1 - \cos k\,(2\,D(x) + x\theta) \right];\ n = 1; \tag{16}$$

where θ is so small that one can set $\sin\theta = \theta$. Since $\theta = 2\alpha$,

$$W = a^2 \left[1 - \cos 2k\,(D(x) + x\alpha) \right] = 2a^2 \sin^2 \left[\frac{2\pi}{\lambda}\,(D(x) + x\alpha) \right] \tag{17}$$

The exact physical significance of D(x) is now clear. Since $x\alpha = d$, the separation between surfaces S_1 and S_2 (see Figure 16.2) at point x, D(x) must be the increase in separation due to a local bulge in one of the reflecting surfaces. $D(x) > 0$ when the bulge increases the separation between the two surfaces.

16.2.2.4 It should be observed from Equation (17) that W = constant whenever

$$D(x) + x\alpha = D(x) + d = \text{constant}. \tag{18}$$

Since $D(x) + d$ is the actual separation of the surfaces at point x, it follows that an interference fringe is the locus of those points x for which the separation of the interferometer surfaces is a constant. If the surfaces are plane, $D(x) = 0$ and the fringes are straight.

16.2.2.5 Suppose one of the interferometer flats is pressed or moved so as to decrease d by a small amount. Since each fringe is the locus of equal separations $D(x) + d$, the whole family of fringes will move in the positive x - direction of Figure 16.2 wherever $D(x) = 0$. In localities where $D(x) = 0$, each fringe will move in a slightly more complex manner so as to find the location where $D(x) + d$ remains constant.

16.3 THE TWYMAN-GREEN INTERFEROMETER

16.3.1 Principles of operation.

16.3.1.1 The essential characteristics of the Twyman Green interferometer are shown in Figure 16.3. The physical principles utilized in the Twyman-Green interferometer and in the Fizeau interferoscope are so similar that the corresponding elements of Figures 16.2 and 16.3 are recognized easily. These corresponding elements are denoted by the same symbols. A small pinhole H, illuminated by monochromatic light, is located at the first focal plane of the collimator L_1 so that a plane wave front is reflected by surfaces S_1 and S_2 of the end-mirrors. A telescope is added to produce an image of the pinhole H at H_1. The surface S_1 appears to be located at S_1'. If S_1' makes the angle α with S_2, the ray reflected from S_1 will appear to be a ray BR such that BR makes the angle $\theta = 2\alpha$ with the ray AQ reflected from S_2. We take OZ parallel to AQ and OP parallel to BR. The coordinate OX falls in the wave front reflected by S_2. This time, to complement Figure 16.2, we show the passage of ray BR to the vicinity of the eye lens where a second image H_2 of the pinhole H is formed. The width of the interference fringes is increased by decreasing the separation $H_1\,H_2$ of the images of the pinhole by tilting mirror S_2 in the direction for decreasing angles θ and α.

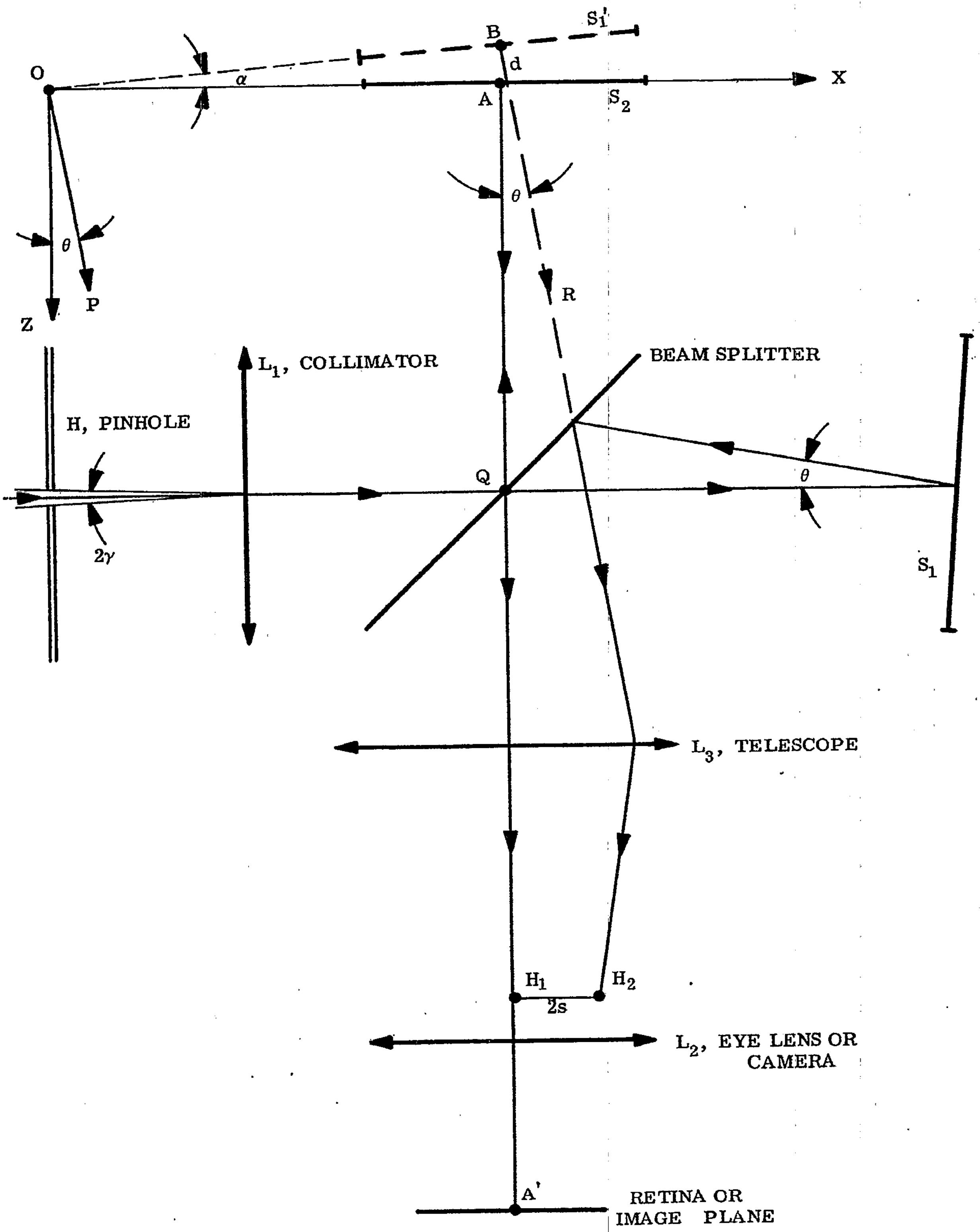


FIGURE 16.3 -Notation with respect to the Twyman Green Interferometer.

16.3.1.2 One major use of the Twyman Green interferometer is to examine the optical quality of glass plates, prisms, etc. The sample to be examined is placed in one arm of the interferometer and the effect of the sample upon the fringes noted. It is usually necessary to readjust the angular setting of at least one of the mirrors S_1 and S_2 and to alter the length arm AQ so as to obtain best contrast in the fringes. We shall not be concerned here with the details of the many applications of the Twyman Green interferometer but, rather, with the principles involved.

16.3.1.3 The beam splitter is usually an optically parallel plate, one of whose surfaces is coated with a uniform film of silver or aluminum. Whereas it is not necessary that the transmittance and reflectance of the filmed surface shall be alike, they should not be markedly dissimilar. Where utmost contrast in the fringes is desired, the second surface of the beam splitter should be rendered low reflecting. Henceforth, it will be supposed that the beam splitter consists, in effect, of a single surface as in Figure 16.3.

16.3.1.4 The amplitudes a_1 and a_2 of the two, plane, interfering waves that enter the telescope L_3 are not likely to be as nearly equal as in the Fizeau interferoscope. However, these amplitudes will be nearly alike provided that the end mirrors S_1 and S_2 have practically equal reflectances and provided that the test sample transmits well. It is, of course, possible to compensate the effects of the sample in one arm by placing a suitable absorbing plate in the second arm.

16.3.1.5 We saw that the phase difference δ between the two interfering waves will be π in the Fizeau interferoscope. The phase changes on reflection at the surfaces S_1 and S_2 of the Twyman Green interferometer are likely to be nearly alike to that δ can be sensibly zero. However, one cannot always be certain that δ is sensibly zero or that a_1 and a_2 are sensibly alike.

16.3.1.6 The product knz $(1 - \cos\theta)$ of Equation (7) will usually be negligible in the Twyman Green interferometer for the same reason that applies to the Fizeau interferoscope. We introduce this approximation into Equation (7). Instead of writing δ as in Equation (15), we set $\delta = \delta_o - 2kn\, D(x)$. The result is

$$2W = a_1^2 + a_2^2 + 2a_1\, a_2 \cos\left[\delta_o - 2k\,(x\alpha + D(x))\right] \qquad (19)$$

since $\sin\theta \rightarrow \theta$ and $\theta = 2\alpha$.

Equations (17) and (19) differ mainly in that the fringe system is shifted slightly with respect to x and in that the fringe contrast obtained from (19) will be inferior to the constrast obtained from (17) except when $a_1 = a_2 = a$, i.e., except when the amplitudes of the two interfering waves are made substantially alike in the Twyman Green interferometer.

16.4 EFFECT OF MONOCHROMATICITY ON FRINGE CONTRAST

16.4.1 Discussion of problem.

16.4.1.1 Fringes obtained with Fizeau interferoscopes or with Twyman Green interferometers "wash out" when the path difference d, Figures 16.2 and 16.3, becomes too great relative to the spectral purity of the monochromatic source. It can be shown that the effect of the presence of many different wavelengths $(k = 2\pi/\lambda)$ in Equation (19) is to reduce the average value of the cosine term to zero as the spread of wavelengths is increased. The physical circumstances become similar to those which cause Equation (3) to degenerate to Equation (4). We may say that the source of light becomes incoherent. An insight into the nature and magnitude of the difficulty can be obtained from the following simplified considerations.

16.4.1.2 With respect to Equation (19), suppose for convenience of argument that $D(x) = 0$ and suppose that the source contains wavelengths from $\lambda_o - |\Delta\lambda|$ to $\lambda_o + |\Delta\lambda|$ where $|\Delta\lambda|$ is small. If $|\Delta\lambda|$ is too large, an interference maximum for $\lambda = \lambda_o$ will fall at the same point $x\alpha = d$ as the first interference minimum due to the wavelength $\lambda = \lambda_o - |\Delta\lambda|$. Thus, if

$$\frac{4\pi x\alpha}{\lambda_o} - \delta_o = \nu\ 2\pi\ ;\ \nu \text{ an integer;} \tag{20}$$

and

$$\frac{4\pi x\alpha}{\lambda_o - |\Delta\lambda|} - \delta_o = \nu\ 2\pi + \pi\ ; \tag{21}$$

then since $1/(\lambda_o - |\Delta\lambda|) \approx 1/\lambda_o + |\Delta\lambda|/\lambda_o^2$, it follows by subtraction of Equation (20) from (21) that $4x\alpha\,|\Delta\lambda|\,/\lambda_o^2 = 1$. When, therefore,

$$|\Delta\lambda| = \lambda_o^2/4x\alpha = \frac{\lambda_o^2}{4d}\ , \tag{22}$$

a bright fringe due to λ_o will fall upon a dark fringe due to $\lambda = \lambda_o - |\Delta\lambda|$. If the radiant fluxes of the two wavelengths are approximately equal and if the wavelengths λ_o and $\lambda_o - |\Delta\lambda|$ do not differ appreciably in color, the interference fringes will be practically obliterated when $|\Delta\lambda|$ and the separation d are related as in Equation (22).

16.4.1.3 When $|\Delta\lambda|$ is less than that given by Equation (22), one can expect that the fringes will be visible. In fact, we must expect from Equation (22) that the condition for the appearance of interference fringes is

$$|\Delta\lambda|\ d < \frac{\lambda_o^2}{4}\ . \tag{23}$$

Contrast in the fringes is improved by choosing $|\Delta\lambda|$ and the path difference d so that their product is small.

16.5 EFFECT OF PINHOLE SIZE ON CONTRAST

16.5.1 Discussion of problem.

16.5.1.1 As can be expected intuitively, the effect of opening the pinhole H too far is to reduce contrast in the fringes even though the light is so monochromatic that one can set $|\Delta\lambda| = 0$. It can be shown that when the pinhole size cannot be neglected, one obtains, instead of monochromatic law of Equation (19), the result

$$2W_T = a_1^2 + a_2^2 + 2a_1 a_2 \left[2\,\frac{J_1\,(k2\gamma s)}{k2\gamma s} \right] \cos \left[\delta_o - 2k\,(x\alpha + D(x)) \right] \tag{24}$$

in which W_T is the total energy density due to all of the points in the illuminated pinhole H (see Figure 16.3), 2γ is the angle subtended by the pinhole at the collimating lens L_1, $2s$ is the indicated separation of the images H_1 and H_2 of the pinhole and J_1 is a Bessel function of first order and first kind.

16.5.1.2 The function $2J_1(t)/t$ assumes its maximum value of unity at $t = 0$. Therefore, Equation (24) is identical to Equation (19) whenever the angle 2γ subtended by the pinhole at the collimator is so small that one can accept the approximation $2J_1\,(k2\gamma s)/k2\gamma s = 1$. Contrast in the fringes is excellent provided that the amplitudes a_1 and a_2 of the interfering beams are not too unlike. For a given value of 2γ, the fringes should show better contrast as they are broadened, i.e., as the separation s of the two pinhole images is decreased.

16.5.1.3 The function $J_1(t)/t$ has an infinite number of zeros the first of which occurs at $t = 3.8317$. Whenever the product $2\gamma s$ becomes so large that $2\pi\ 2\gamma s/\lambda = 3.8317$, $J_1(k2\gamma s)/k2\gamma s = 0$. Hence W_T becomes constant and should be independent of x and the fringes should vanish when

$$2\gamma s = \frac{3.8317\lambda}{2\pi} = 0.61\lambda\ . \tag{25}$$

Since $J_1(t)/t$ changes sign as t passes through any of the roots of $J_1(t)/t = 0$, the fringes should shift abruptly by one half fringe width as $2\gamma s$ passes through the value given by Equation (25).

16.5.1.4 Whereas it is the writer's experience that Equation (25) does not agree in an excellent quantitative manner with observations in, say, the Twyman-Green interferometer, it does serve as semi-quantitative basis for predicting the degree of contrast in the fringes.

16.6 YOUNG'S PINHOLE INTERFEROMETER

16.6.1 Introduction.

16.6.1.1 The Fizeau and Twyman Green interferometers belong to a broad class of doubled pinhole interferometers in which two actual pinholes are illuminated or in which the image of one illuminated pinhole is doubled by any one of a variety of beam splitting devices. The rudimentary theory of formation of the interference fringes is essentially the same for this group of interferometers. If the plane of observation is sufficiently far from the location of the pinholes, the two corresponding waves that arrive at the plane of observation are essentially plane so that the theory of the foregoing paragraphs applies.

16.6.1.2 The following argument presents a second, very useful point of view that encroaches to some extent upon Huygens' principle. Let us consider the simplest of all double pinhole interferometers, namely Young's famous interferometer of Figure 16.4. Monochromatic light is focused upon a small pinhole H. Coherent, spherical waves emanate from H and illuminate the small pinholes H_1 and H_2. If H falls upon the Z-axis, the light reaching H_1 and H_2 will be in phase. Otherwise, a phase difference δ_o will be introduced. Pairs of coherent, spherical waves emerge from pinholes H_1 and H_2 and reach point (x, y) of the observation plane after traversing paths r_1 and r_2. If distance D is large relative to the separation 2s of the pinholes H_1 and H_2 and if point(x, y) is not too far from the Z-axis, the distances r_1 and r_2 will be so nearly alike that the two waves from H_1 and H_2 will arrive at point (x, y) with substantially equal amplitude provided that they leave H_1 and H_2 with substantially equal amplitude. We shall suppose for sake of generality that the interfering waves from H_1 and H_2 reach point(x, y) with the amplitudes a_1 and a_2, respectively. (The amplitude of one of the waves might be reduced, for example, by placing an absorbing glass plate over one of the pinholes or by making the pinholes small but unlike in area).

From Figure 16.4,

$$r_1^2 = D^2 + (x - s)^2 + y^2 ; \tag{26}$$

$$r_2^2 = D^2 + (x + s)^2 + y^2 . \tag{27}$$

therefore

$$r_2^2 - r_1^2 = (r_2 - r_1)(r_2 + r_1) = 4xs \tag{28}$$

or

$$r_2 - r_1 = 2xs\Big/\left(\frac{r_2 + r_1}{2}\right). \tag{29}$$

It matters to a considerable extent which approximation one wishes to accept for $(r_2 + r_1)/2$, the average value of r_1 and r_2. In case the point of observation,(x, y), Figure 16.4, falls near the Z-axis, both r_1 and r_2 differ only slightly from $R = \sqrt{D^2 + s^2}$, and the average value of r_1 and r_2 will fall nearer R than either r_1 or r_2. Accordingly, we suppose that the point of observation,(x, y), falls near the Z-axis and accept the approximation

$$r_2 - r_1 = \frac{2xs}{\sqrt{D^2 + s^2}} \tag{30}$$

Then from Figure 16.4, $\dfrac{s}{\sqrt{D^2 + s^2}} = \sin\dfrac{\theta}{2}$. Since D is great relative to s,

$$\sin\frac{\theta}{2} = \frac{\theta}{2} = \frac{s}{\sqrt{D^2 + s^2}} ; \tag{31}$$

therefore,

$$r_2 - r_1 = x\theta \tag{32}$$

in which θ is very nearly equal to the actual angle between the direction of propagation of the two waves that reach point(x, y) from the pinholes H_1 and H_2.

16.6.1.3 We find that the two coherent waves which interfere at point (x, y) have amplitudes a_1 and a_2 and the phase difference $\phi_1 - \phi_2$ such that

$$\phi_1 - \phi_2 = \delta_o + k(r_1 - r_2) = \delta - kx\theta \tag{33}$$

wherein the portion $kx\theta$ is due to the path difference $r_2 - r_1$ and wherein δ_o specifies the phase difference

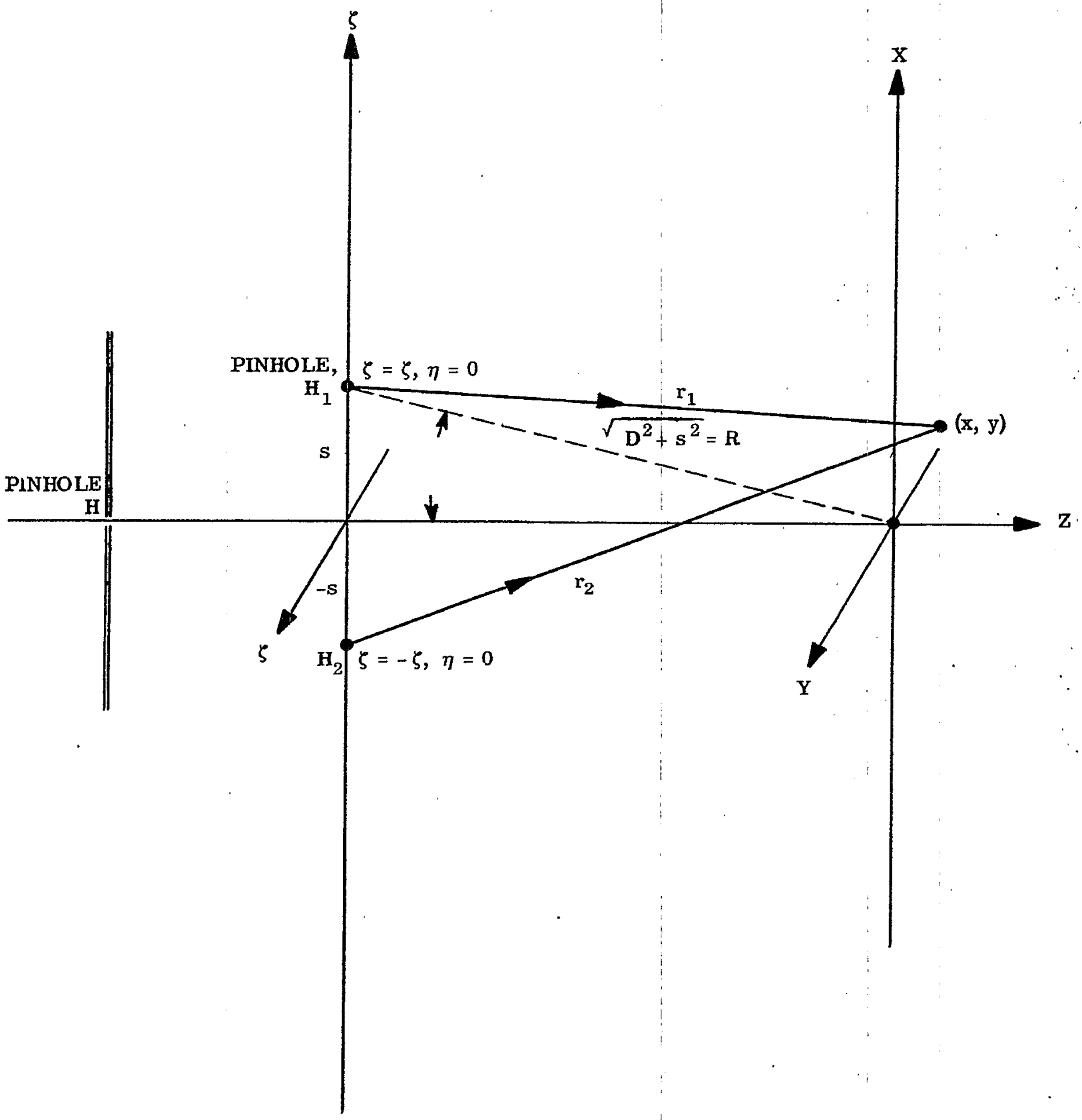


FIGURE 16.4 - Young's Pinhole Interferometer

between the two, interfering, non-collinear waves as they leave the pinholes H_1 and H_2. The time-averaged energy density W produced by interfering waves is given again by Equation (3). Thus, from Equations (3) and (33)

$$2W = a_1^2 + a_2^2 + 2 a_1 a_2 \cos (\delta_o - kx\theta) \tag{34}$$

in which

$$k = 2\pi/\lambda \quad \text{and} \quad \theta = 2s / \sqrt{D^2 + s^2} \; . \tag{35}$$

where 2s is the separation of the pinholes H_1 and H_2, Figure 16.4.

16.6.1.4 Comparision of Equations (19) and (34) shows that they agree since $2\alpha = \theta$ and since D(x), as defined in Paragraph 16.3, is zero as applied to Young's pinhole interferometer.

16.6.1.5 The fringes formed in Young's pinhole interferometer will not be straight, as predicted by the approximation of Equation (34), unless the point of observation is near the Z-axis of Figure 16.4. As the point of observation is moved out to distances $\sqrt{x^2 + y^2}$ that become appreciable with respect to D, the average value of r_1 and r_2 becomes a function of both x and y. It follows from Equations (29) and (33) that $\phi_1 - \phi_2$ will not vary in a simple linear manner with x. The fringes become curved in a manner that is not difficult to ascertain from a further study of Equations (26), (27), and (29).

16.6.1.6 Young's slit interferometer is obtained by replacing pinholes H, H_1 and H_2 by very narrow slits perpendicular to the plane of the paper. With this arrangement, the interference fringes seen at plane x y will remain straight over a greatly increased portion of the xy plane provided that the slits are sufficiently long.

16.6.1.7 Young's interferometer is both useful and simple to construct. The difference in optical path, for example, of two similar glass plates of nearly the same thickness can be ascertained by applying the following principles. We observe that if pinhole H is on the Z axis, Figure 16.4, a bright white fringe will be formed at O, where x = 0, when H is illuminated with white light because the optical paths from H to O are equal. Constructive interference occurs at O for all wavelengths. If the pinhole H is not on the Z axis, the bright white fringe will be found at a location $x \neq 0$. This location is called the white light position and determines a point of reference at which the optical paths from H to O are equal. When monochromatic light is substituted for white light, the fringes appear in best contrast about the white light position. Suppose that the optical path H H_1 O is increased by a slight amount δ_o relative to the optical path H H_2 O by the insertion at H_1 and H_2 of glass plates that differ slightly in optical path. We see from Figure 16.4 that the ray H_2 x must be inclined toward larger x-values in order to equalize the optical path difference between the paths H H_1 x and H H_2 x. Therefore, the white light fringe or any monochromatic fringe must move outward from the axis Z in the direction of that pinhole H_1 or H_2 over which has been placed the plate having the greater optical path. The magnitude of δ_o can be found as follows from the measurement of the fringe shift produced by δ_o.

16.6.1.8 First, the fringe width h is the increase in x for which $k(x + h)\theta$ exceeds $kx\theta$ by 2π in Equation (34). Thus

$$kh\theta = 2\pi \quad \text{or} \quad h = \frac{\lambda}{\theta}. \tag{36}$$

Secondly, a given interference fringe occupies that position x for which

$$\delta_o - kx\theta = \text{constant} = C. \tag{37}$$

Suppose, for generality, that δ_o has successively the values δ_1 and δ_2. Denote the corresponding position of a given fringe by x_1 and x_2. Then from Equation (37)

$$\begin{aligned} \delta_1 - kx_1\theta &= C \\ \delta_2 - kx_2\theta &= C \end{aligned} \tag{38}$$

By subtraction of Equations (38) one finds that

$$\delta_2 - \delta_1 = k\theta (x_2 - x_1) = \frac{2\pi}{\lambda} \theta (x_2 - x_1). \tag{39}$$

From Equations (37) and (39) we obtain the extremely useful result

$$\delta_2 - \delta_1 = 2\pi \frac{x_2 - x_1}{h}, \text{ radians.} \tag{40}$$

In other words, the phase change in the two arms H H_1 x and H H_2 x is given by the ratio of the fringe shift, $(x_2 - x_1)$, to the fringe width, h.

16.6.1.9 Difficulties can appear when $\delta_2 - \delta_1$ exceeds 2π; for then the fringe shift, $x_2 - x_1$, exceeds the fringe width, h, by a number of fringes that may not be obvious. This ambiguity about the "fringe jump" can be settled by considering the shift of the white light position or by making measurements of the fringe locations at more than one wavelength.

16.7 LLOYD'S INTERFEROMETER

16.7.1 Description. Lloyd's double pinhole or double slit arrangement for obtaining interference fringes is illustrated in Figure 16.5. Corresponding elements are denoted by the same symbols in Figures 16.4 and 16.5 to emphasize their similarity. The interpretations of the interferometers due to Lloyd, Young, Fizeau, and Twyman Green are alike provided that the pinholes are small and provided that the distance D is great. It should be observed that the virtual image H_2 is a mirror image of H_1. The relative locations of the corresponding coherent points in the "images" H_1 and H_2 will therefore be significantly different in Lloyd's interferometer as compared with the Fizeau and Twyman Green interferometers. This mirror image relation between H_1 and H_2 is avoided by Fresnel's double mirror interferometer which is illustrated in Figure 16.6. Both H_1 and H_2 are now virtual images whose separation 2s is governed by the angle α between the interferometer mirrors M_1 and M_2.

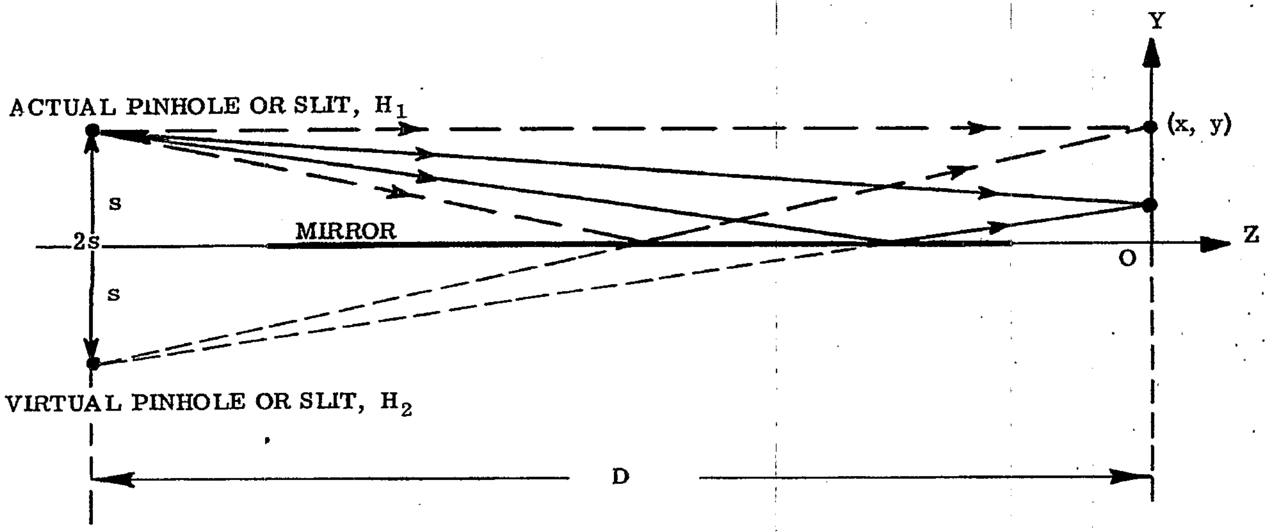


FIGURE 16.5 - Lloyd's Interferometer

16.8 FRESNEL COEFFICIENTS FOR NORMAL INCIDENCE

16.8.1 Computing amplitude reflectance and transmittance.

16.8.1.1 Let n and nK denote the optical constants of two media that are in contact across a plane interface as in Figure 16.7. Then for normal incidence upon the interface along the indicated direction, the amplitude reflectance ρ is given by

$$\rho = \frac{M_0 - M_1}{M_0 + M_1} \tag{41}$$

and the amplitude transmittance τ across the interface is given by

$$\tau = \frac{2 M_0}{M_0 + M_1} \tag{42}$$

where

$$M_\nu \equiv n_\nu (1 + i K_\nu); \quad \nu = 0, 1. \tag{43}$$

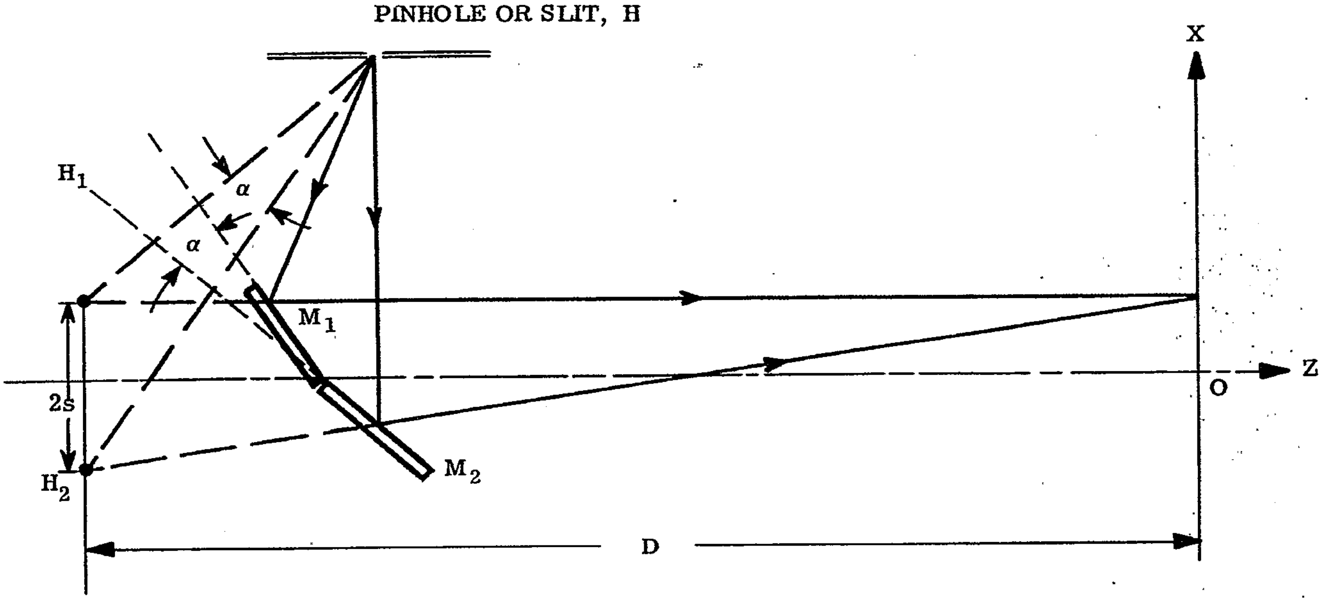


FIGURE 16.6 - Fresnel's Mirror Interferometer.

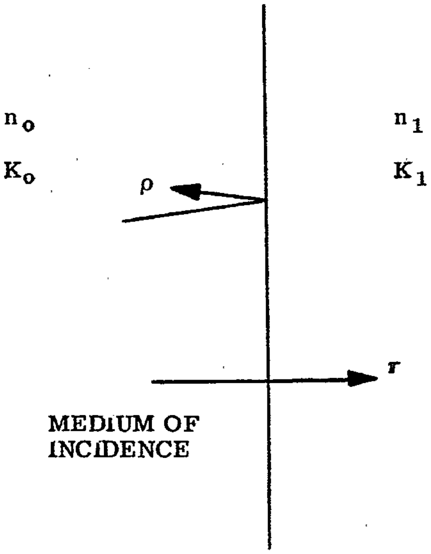


FIGURE 16.7 - Transmittance and reflectance at Translucent Interface

16.8.1.2 Suppose that neither medium absorbs so that $K_o = K_1 = 0$. Then

$$\rho = \frac{n_o - n_1}{n_o + n_1}; \qquad \tau = \frac{2n_o}{n_o + n_1}. \tag{44}$$

We see that $\tau > 0$, but the amplitude reflectance ρ is greater or less than zero according as n_o is greater than or less than n_1. If we write ρ in the form

$$\rho = \left| \frac{n_o - n_1}{n_o + n_1} \right| \cos \sigma, \tag{45}$$

we see that σ, the phase change on reflection, is zero when $n_o > n_1$ but is π when $n_o < n_1$. Furthermore, the phase change on transmission across an interface between two non-absorbing media is always zero.

16.8.1.3 Interferometers usually involve the splitting of a light beam at one or more interfaces between two media. In order to compute or to estimate the amplitudes a_1 and a_2 of the interfering waves thus produced, knowledge of the Fresnel coefficients is essential. The Fresnel coefficients at normal incidence will suffice for the purposes of the present text. The application of Fresnel's coefficients for normal incidence to cases involving oblique incidence can, however, be misleading. The reader who needs to compute the amplitudes a_1 and a_2 for oblique incidence should consult paragraph 24.1.

16.9 INTERFERENCE WITH PLANE PARALLEL PLATES AND DISTANT LIGHT SOURCES

16.9.1 Discussion of Problem.

16.9.1.1 A ray AB from one point in a distant source of light is incident upon a plate of thickness d with refractive index n_1. Reflected rays R_1, R_2, R_3, etc., and transmitted rays T_1, T_2, etc., are formed in the manner indicated in Figure 16.8. We suppose that the plate is non-absorbing and that the reflectance of its surfaces is so low that only rays R_1 and R_2 need be considered in the reflected beam of rays. The problem is to find the optical path difference δ between rays R_2 and R_1 under the assumption that the surfaces of

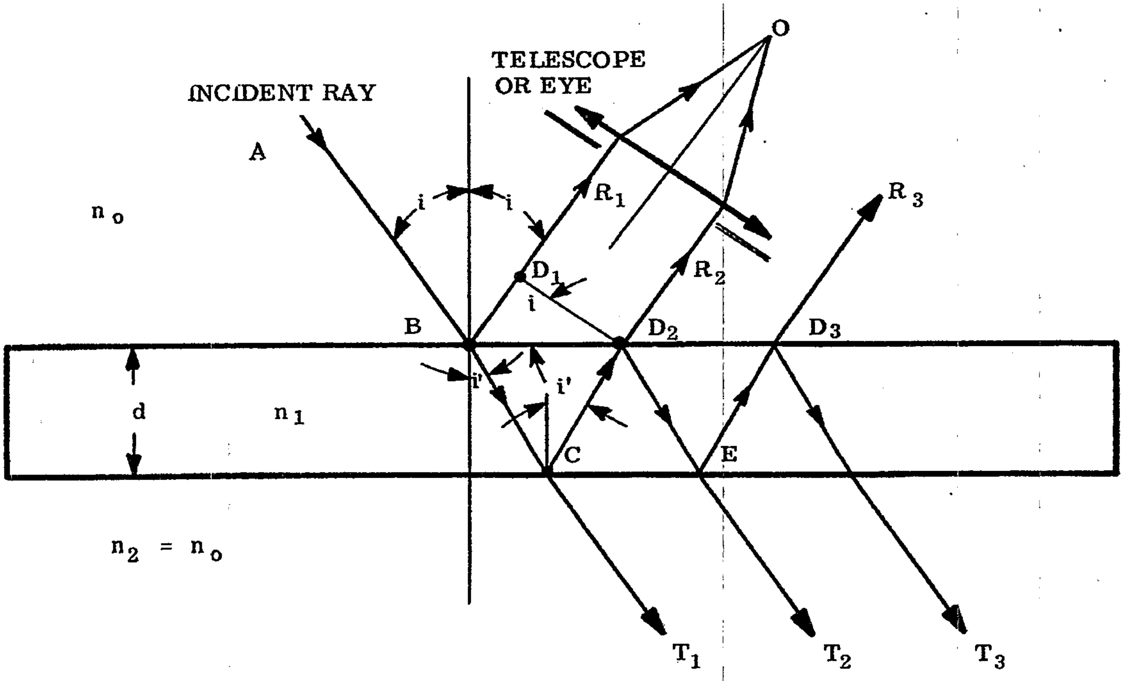


FIGURE 16.8 - The Dielectric Plate as an Interferometer.

the plate are parallel. Let line segment $D_1 D_2$ be drawn perpendicular to rays R_1 and R_2. Then

$$\delta = n_1 (BC + CD_2) - n_0 BD_1 \tag{46}$$

$$BC = CD_2 = \frac{d}{\cos i'}$$

$$BD_2 = 2d \tan i'$$

$$BD_1 = BD_2 \sin i = 2d \tan i' \sin i. \tag{47}$$

Substitution of relations (47) into Equation (46) yields

$$\delta = \frac{2n_1 d}{\cos i'} \left(1 - \frac{n_0}{n_1} \sin i \sin i' \right) . \tag{48}$$

Since $n_0 \sin i = n_1 \sin i'$, it follows that the optical path difference δ between rays R_1 and R_2, Figure 16.8, is given by

$$\delta = 2n_1 d \cos i' \tag{49}$$

where n_1 and d are, respectively, the refractive index and thickness of the plate, and i' is the indicated angle of refraction. Equation (49) is of great importance to the interpretation of interferometry with films and plates.

16.9.1.2 Let us suppose that the plate is immersed in a single medium. Then $n_2 = n_0$. It follows from the principles of the preceding section that the phase changes on reflection at B and C, Figure 16.8, differ by π radians. Thus,

$$\Delta = \frac{2\pi}{\lambda} 2n_1 d \cos i' + \pi \text{ radians} \tag{50}$$

where Δ is the total phase difference introduced between rays R_1 and R_2 due to the optical path difference δ and the phase changes on reflection. We have supposed tacitly that the angle i' is not so large that it is essential to distinguish sharply between normal and oblique incidence.

16.9.1.3 The optical path difference between the transmitted rays T_1 and T_2 is also given by δ as in Equation (49). More generally, the optical path difference between any two, consecutive reflected or transmitted rays, such as R_2 and R_3, is given by δ.

16.9.1.4 According to paragraph 16.8, the amplitude a of the wave reflected at points B in the first surface,, Figure 16.8 will be

$$a_1 = \frac{|n_0 - n_1|}{n_0 + n_1} \tag{51}$$

The wave corresponding to rays R_2 is transmitted twice through the first surface in opposite directions and is reflected at points C. Hence, from Equations (44),

$$a_2 = \frac{2n_0}{n_0 + n_1} \frac{|n_1 - n_0|}{n_1 + n_0} \frac{2n_1}{n_0 + n_1} = \frac{4n_0 n_1 |n_1 - n_0|}{(n_0 + n_1)^3} . \tag{52}$$

If, for example, $n_0 = 1$ and $n_1 = 1.5$, then $a_1 = 0.2$ and $a_2 = 0.192$ so that a_1 and a_2 are substantially alike. These two collinear waves interfere to produce the time-averaged energy density or illumination at O, Figure 16.8, proportional to W as given by Equation (3) with δ. From Equations (3) and (50)

$$2W = a_1^2 - 2 a_1 a_2 \cos \left(\frac{4\pi n_1 d}{\lambda} \cos i' \right) + a_2^2 . \tag{53}$$

The illumination produced by interference in the reflected beam can therefore be varied by changing any one of the following parameters:

(a) The optical thickness end of the plate
(b) The angle of refraction, i'
(c) The wavelength, λ.

The illumination at point O, Figure 16.8, is minimum when

$$4\pi n_1 d \frac{\cos i'}{\lambda} = \nu\, 2\pi; \quad \nu = 0, 1, 2, 3, \text{ etc.} \tag{54}$$

On the other hand, this illumination is maximum when

$$4\pi n_1 d \frac{\cos i'}{\lambda} = \mu\, \pi; \quad \mu \text{ an odd integer.} \tag{55}$$

The minima will be quite dark since a_1 and a_2 are substantially alike.

16.9.1.5 It is emphasized that with distant sources of light, the eye or telescope is focused for infinity, as illustrated in Figure 16.8, in order to observe the phenomena discussed in this section.

16.10 INTERFERENCE WITH PLANE PARALLEL PLATES AND NEARBY LIGHT SOURCES

16.10.1 Discussion of Problem.

16.10.1.1 The manner in which interference phenomena can be observed with nearby light sources is illustrated in Figure 16.9. Consider the coherent spherical wave that enamates from point S in the source. Suppose that the eye or camera is focused upon the upper surface of the plate and that the distances SD_2 and SB are large compared to the thickness d of the plate. A pair of rays SBR_2 and SD_2CBR_1 leaves point S and reaches point O in the manner indicated.

16.10.1.2 With point S as center and SD_2 as radius, draw arc D_2D_1. If the distance SD_2 is large and if the thickness d is relatively small, the arc D_2D_1 will be practically straight and perpendicular to SB. Moreover, the difference between angles i_1 and i_2 will be so small that either i_1 or i_2 or an intermediate angle, such as i, can be regarded as the angle of incidence together with i' as the angle of refraction. The optical path difference δ between rays SD_2CBO and SBO is

$$\delta = n_1 (BC + CD_2) - n_o BD_1. \tag{56}$$

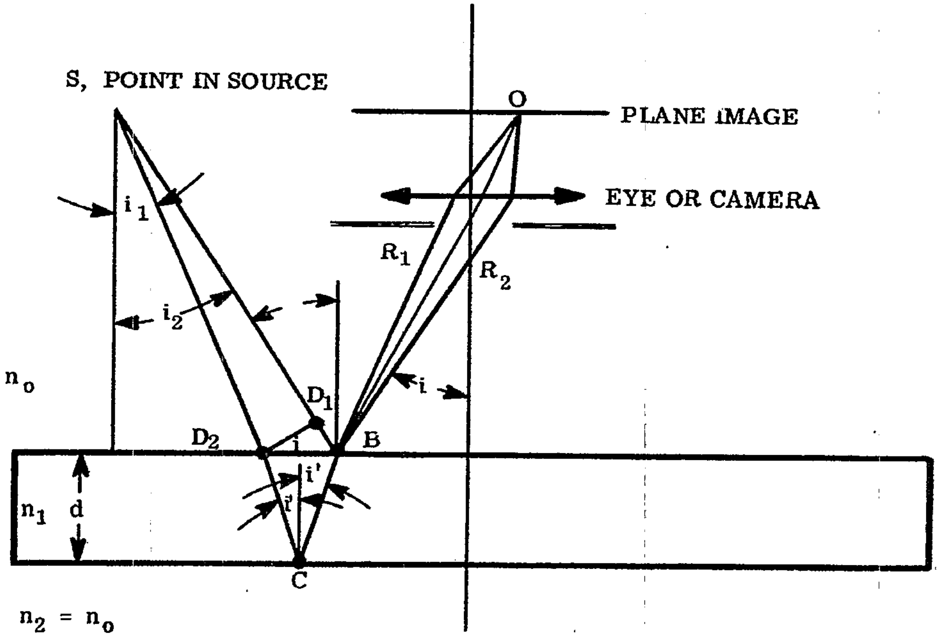


FIGURE 16.9 - The Parallel Plate Interferometer with nearby light sources.

Comparison of Equations (56) and (46) shows that they are alike. Moreover, comparison of points B, C, D_2, and D_1 in Figures 16.9 and 16.8 shows that they play similar roles. Hence

$$\delta = 2n_1 d \cos i' \tag{57}$$

as in Equation (49), and what has been said in the preceding section applies with excellent approximation to illumination with nearby sources provided that the thickness d of the plate is small as compared to the distance from the plate to the source.

16.11 HAIDINGER'S INTERFERENCE FRINGES

16.11.1 Interpretation of Haidinger's Fringes.

16.11.1.1 A simple arrangement for observing Haidinger's fringes is shown in Figure 16.10. The eye is preferably focused at infinity, where fringe contrast is best, but can be focused when desired on any suitable plane B.

16.11.1.2 The discussions in paragraphs 16.9 and 16.10 apply directly to the interpretation of Haidinger's fringes. In the interests of simplicity, let us accept the approximation $a_1 = a_2 = a$ in writing the energy density W of Equation (53) so that

$$2W = a^2 \left[2 - 2 \cos \left(\frac{4\pi n_1 d \cos i'}{\lambda} \right) \right] .$$

Therefore, the energy density in the observed fringes is proportional to

$$W = a^2 \left[1 - \cos \left(\frac{4\pi n_1 d \cos i'}{\lambda} \right) \right], \tag{58}$$

with

$$\sin i = n_1 \sin i' . \tag{59}$$

Dark fringes or bright fringes are seen at angles i, Figure 16.10, for which cos i' obeys Equations (54) or (55), respectively. Since the angles i or i' are constant on circles about the axis AO, Haidinger's fringes are observed as circular fringes about an axis AO that moves with the observer's eye.

16.11.1.3 Suppose, for example, that $n_1 = 1.5$, $d = 1.8$mm, and $\lambda = 0.54 \times 10^{-3}$ mm. Then $2n_1 d/\lambda = 10^4$. Therefore, from Equation (54), $\nu = 10^4$ when $i' = 0$. The order number $\nu = 10^4$ is the highest possible order - and for it the central fringe is black. The next black fringe occurs when $\nu = 9999$, i.e., when

$$\cos i' = \frac{9999}{\frac{2n_1 d}{\lambda}} = 0.9999 \quad \text{or} \quad i' = 0.81^\circ .$$

Since $\sin i = n_1 \sin i'$, the angle i subtended at the observer by the radius of the first dark ring is 1.21°. Because the angular resolving power of the eye is approximately one minute of arc, plates much thicker than 1.8mm can be inspected for parallelism with the unaided eye by moving the plate along the arrow direction Q of Figure 16.10.

16.11.1.4 In applying Haidinger's fringes to the inspection of parallelism of plates, the distance from the eye to the plate should be made three feet, or so. The point O, Figure 16.10, can then be in the plate itself, i.e., one can focus his eye approximately upon the plate. Even though the plate may not be plane parallel, substantially circular Haidinger's fringes will be seen. The central Haidinger fringe will oscillate in brightness a number of times that depend upon the departure of the surfaces of the plate from parallelism as the plate is moved across the field of view in the Q-direction of Figure 16.10. At the central fringe, $\cos i' = 1$. Equation (54) now shows that when ν changes by unity (that is, as the central fringe changes from one state of blackness to the next), the corresponding change Δd in thickness is given by $2n_1 \Delta d/\lambda = 1$ or by

$$\frac{n_1 \Delta d}{\lambda} = \frac{1}{2} . \tag{60}$$

This means that each time the central fringe passes through one cycle, the optical path through the plate has changed (as could be expected without the aid of the theory) by one-half wavelength. Counting the number of blinks of the central Haidinger fringe forms a sensitive and simple method for measuring the amount of departure from parallelism of a plate.

16.11.1.5 It is worth noting that Haidinger's fringes are essentially fringes of equal inclination. Each fringe corresponds to a definite angle i' of inclination. Changes in i' (rather than changes in λ or n_1 d) govern the observed changes in the fringes.

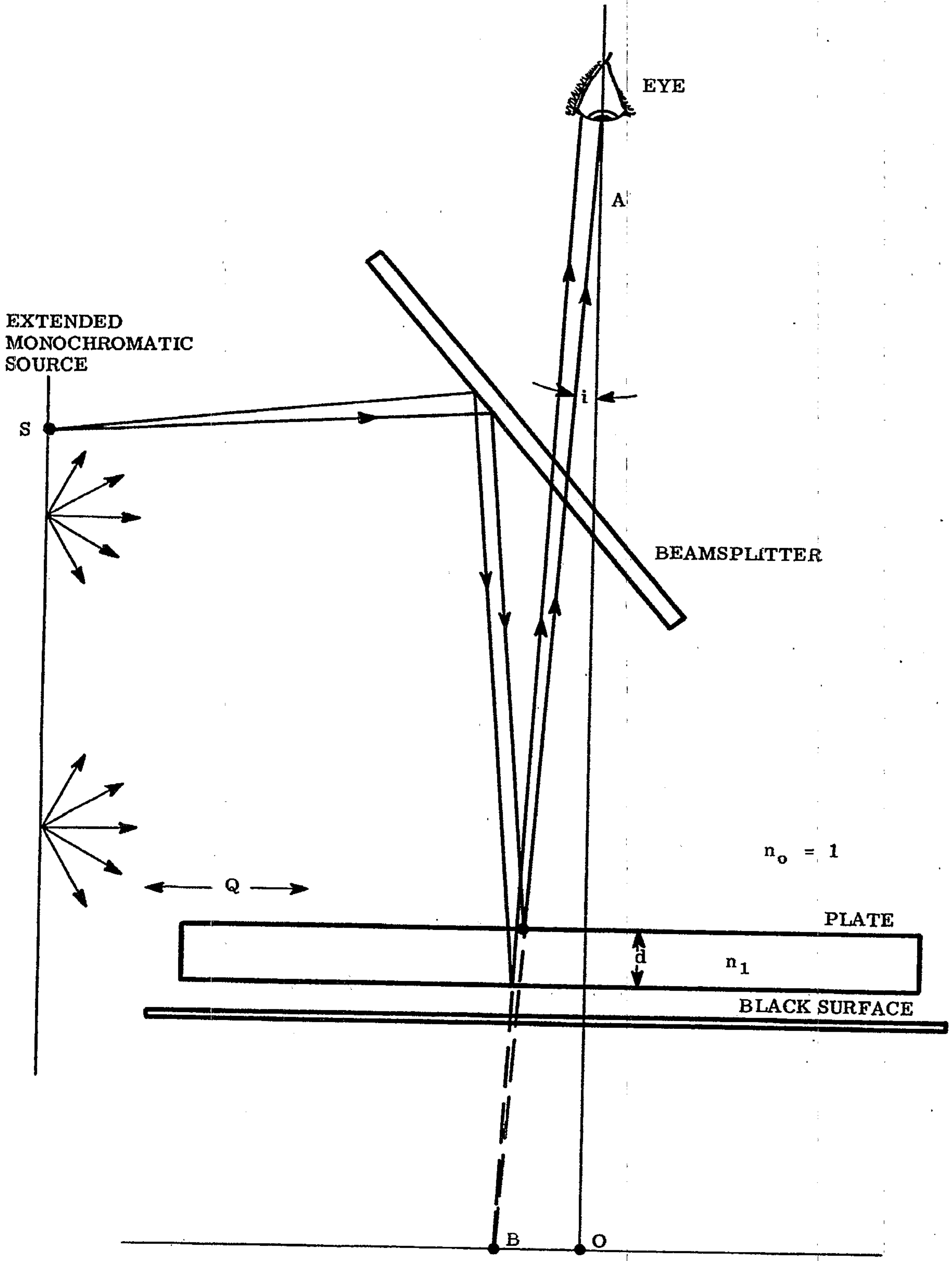


FIGURE 16. 10 -Arrangement for observing Haidinger's Fringes.

16.12 FIZEAU FRINGES

16.12.1 Introduction.

16.12.1.1 The fringes seen with the arrangement shown in Figure 16.11 under illumination from an extended and fairly monochromatic source are often called Fizeau fringes or Fizeau bands. These fringes are similar in formation to those obtained in Michelson's interferometer of Figure 16.12. The method of Figure 16.11 is used widely for testing one polished surface against another for flatness or for sphericity. The reference surface, S_2, may be flat or spherical.

16.12.1.2 Owing to the presence of dust, surfaces S_1 and S_2 will ordinarily be inclined so that the space between them is approximated by an air wedge whose angle θ is constant only when both surfaces are plane. Figure 16.13 illustrates how the Fizeau fringes can appear localized in a chosen plane containing point P. Note that each point P receives coherent light from a corresponding point S in the source. Each point P is, in effect, illuminated by a different point in the source. An extended source becomes necessary for viewing fringes over an extended surface S_1.

16.12.1.3 It will be observed that Figures 16.13 and 16.9 are so similar that they become identical when $\theta = 0$. The argument leading to Equation (57) for the optical path difference δ between rays SPP' and SQPP' applies again with excellent approximation provided that one takes for d the thickness of the wedge at point P as indicated in Figure 16.13. Equations (54) and (55) govern the location of the fringes. Minima occur where

$$2n_1 \, d \cos i' = \nu \lambda ; \quad \nu = 0, 1, 2, 3, \text{ etc.}, \tag{61}$$

and maxima occur where

$$4n_1 \, d \cos i' = \mu\lambda ; \quad \mu \text{ an odd integer.} \tag{62}$$

Since the space between S_1 and S_2 is usually air,

$$n_1 = 1$$
$$i' = i \tag{63}$$

where i is the angle of incidence.

16.12.1.4 The advantage of simplicity obtained through the use of Fizeau fringes rests upon the fact that variations in the angle of incidence i, Figure 16.11, have negligibly small effects upon the location of the fringes because the separations d between S_1 and S_2 are small. Suppose, for example, that $d = 10\lambda$. The maximum value of ν occurs at $i' = 0$, and here $\nu = 20$ from Equation (61) since $n_1 = 1$. If d and λ remain constant as point P moves away from point O, the next dark fringe occurs at $\nu = 19$ so that $\cos i' = \cos i = 19/20$. Correspondingly, $i = 18.19^{\circ}$. Let the distance AO from the eye to the test plate be made so large relative to the lateral dimension of the test plate that the maximum value of i cannot exceed 4°. The variation of 4° is obviously small compared to the amount 18.19° required for decreasing $d \cos i'$ by the amount $\lambda/2$. In fact when $0 \leqq i' \leqq 4^{\circ}$, $0.9976 \leqq \cos i' \leqq 1$. Hence, $10\lambda \geqq d \cos i' \geqq 9.976\lambda$. This means that $d \cos i'$ cannot change by more than 0.034 wavelengths due to any variation of the angle of incidence when i_{max} is constrained to 4° by the choice of the distance AO. If, therefore, one arranges to observe the Fizeau fringes at normal incidence, he is justified in setting $\cos i' = 1$ in Equations (61) and (62) and accepting the well known approximation that the separation d changes by the amount $\lambda/2$ in passing, for example, from one bright fringe to the next. Each fringe may be regarded as the locus of points for which the separation of the surfaces S_1 and S_2, Figure 16.11, is constant.

16.13 NEWTON'S RINGS AND NEWTON'S FRINGES

16.13.1 Interpretation of Newton's Fringes.

16.13.1.1 An experimental arrangement for observing Newton's rings or fringes is illustrated in Figure 16.14 In honor of Sir Isaac Newton, the colored circular fringes seen around the point O with white light sources are called Newton's rings. The central fringe is black at O when S_1 and S_2 are substantially in contact because there is a phase difference of one-half vibration between the reflections at S_1 and S_2. It is preferable for most purposes to view the interference bands with monochromatic sources. These circular bands are often called Newton's Fringes. Comparison of Figures 16.11 and 16.14 shows that Fizeau's and Newton's fringes can become practically identical.

16.13.1.2 It was seen in the previous section that a Fizeau fringe can be regarded with good approximation as the locus of points for which the separation of the surfaces S_1 and S_2 is constant. We may take the view that

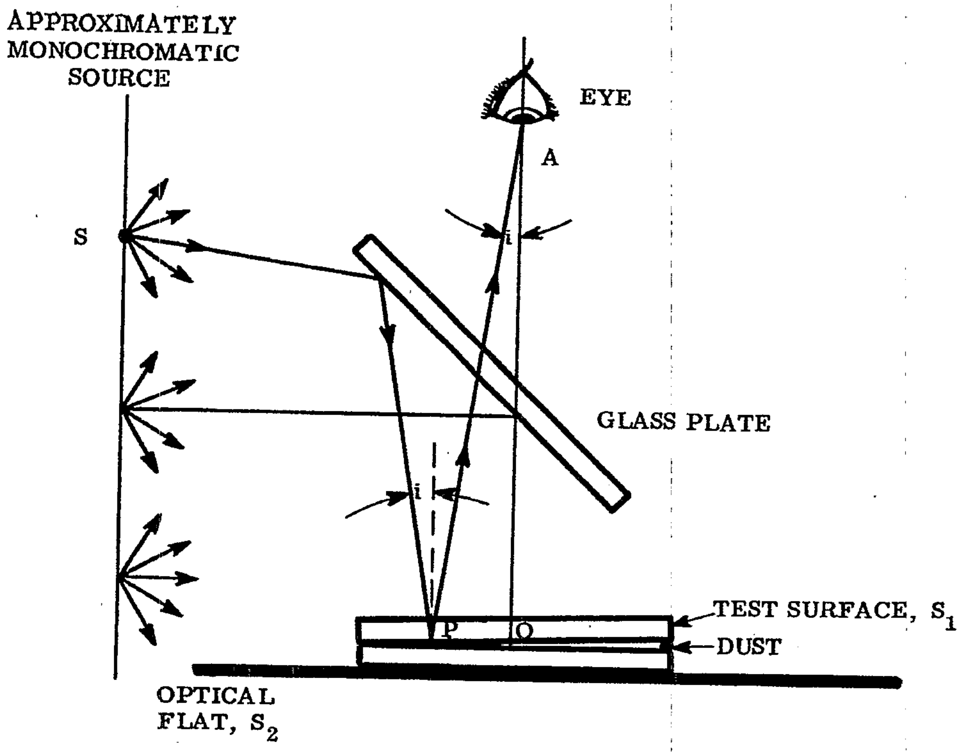


FIGURE 16. 11- Method for obtaining Fizeau Fringes.

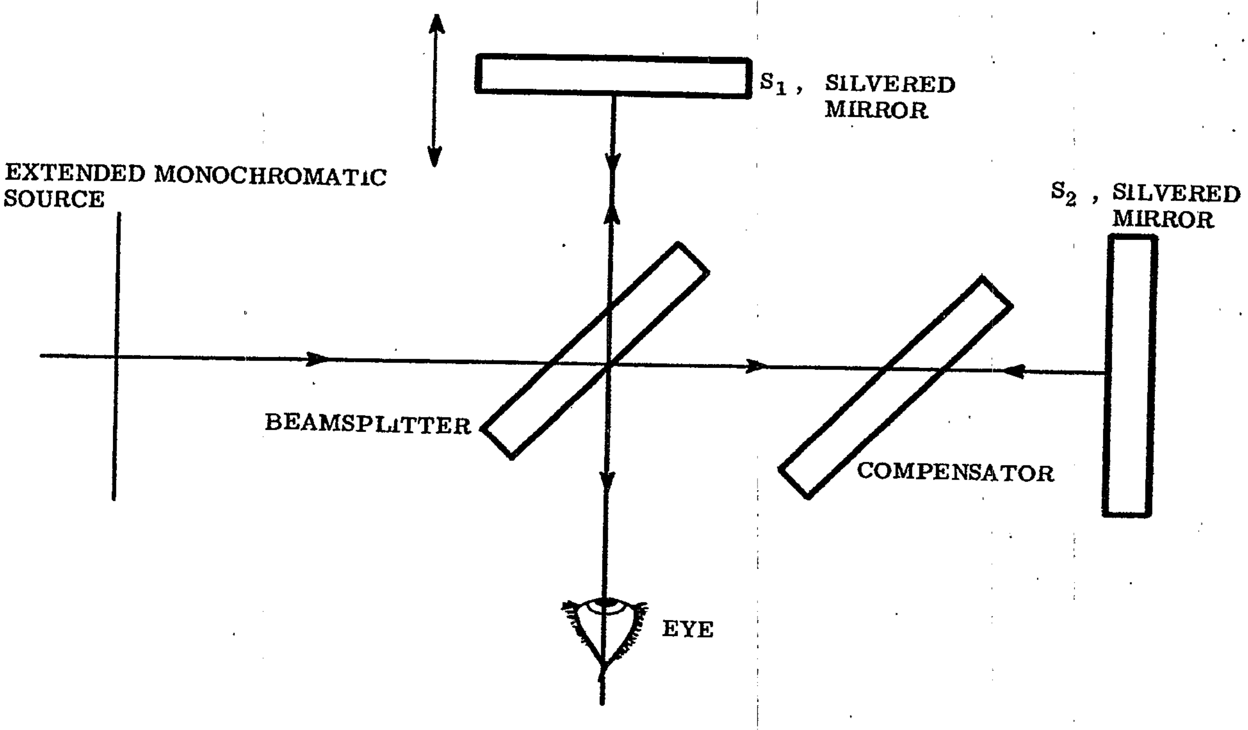


FIGURE 16. 12-Michelson's Interferometer.

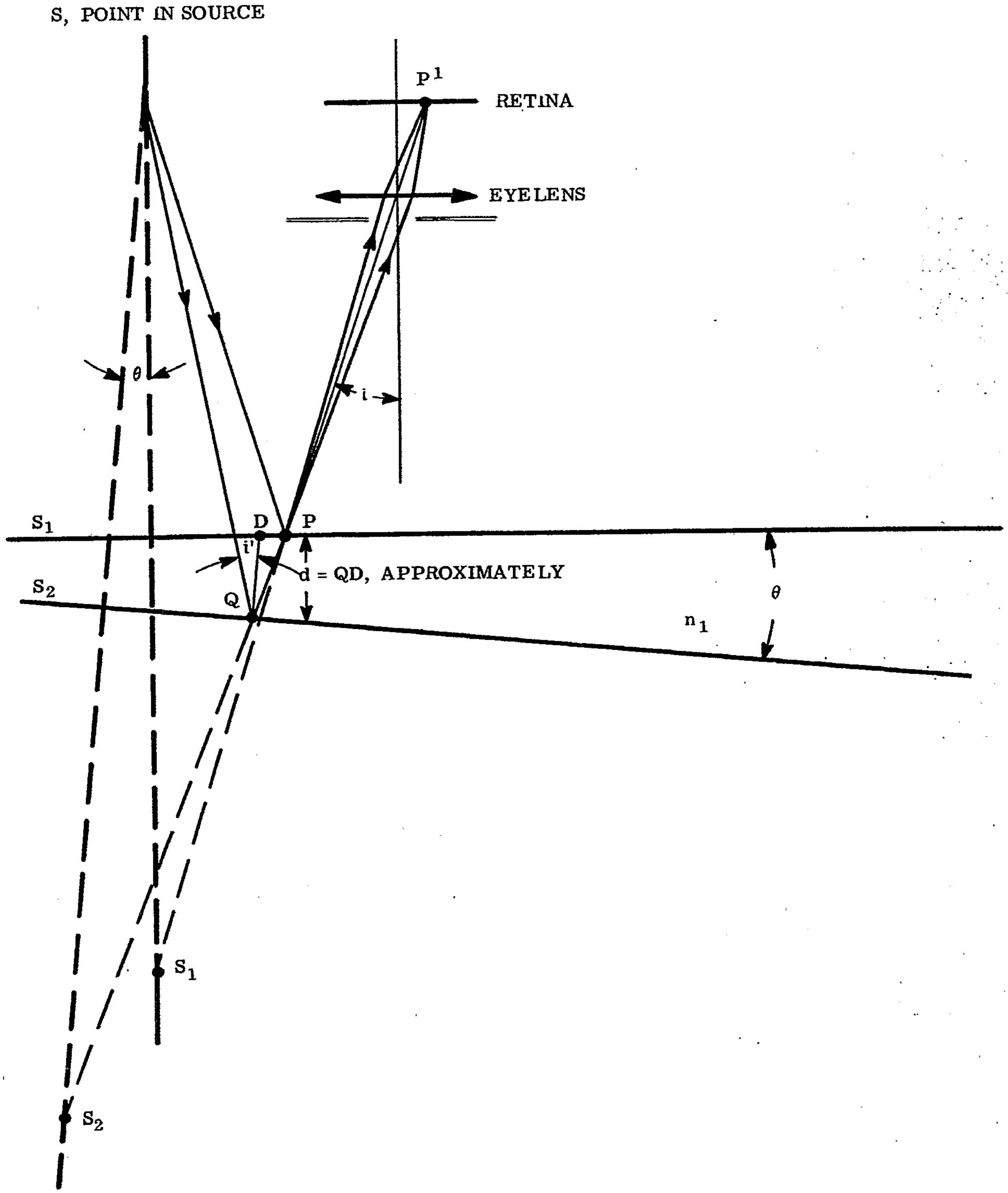


FIGURE 16. 13 - Construction showing how the Fizeau Fringes can appear localized at points, P, near the reflecting surfaces S_1 and S_2

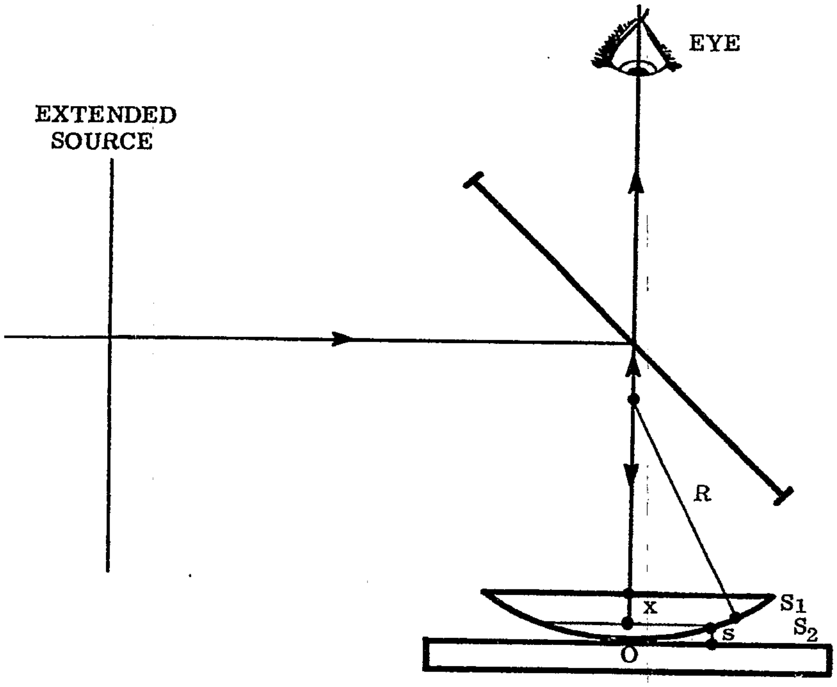


FIGURE 16. 14-Arrangement for obtaining Newton's Rings or Newton's Fringes.

Newton's fringes are Fizeau fringes along which the sagitta s of Figure 16. 14 is constant such that dark fringes occur when

$$s = \nu \frac{\lambda}{2} ; \quad \nu = 0, 1, 2, 3, \text{etc.}, \tag{64}$$

and such that bright fringes occur when

$$s = \mu \frac{\lambda}{4} ; \quad \mu = 1, 3, 5, \text{etc.} \tag{65}$$

The sagitta s obeys the relation $x^2 = 2Rs - s^2$, where R is the radius of the surface. By neglecting s^2 in comparison with 2Rs, one obtains the approximation

$$x = \sqrt{2Rs} . \tag{66}$$

Thus, from Equations (64), (65), and (66)

$$x_\nu = \sqrt{\nu R \lambda} ; \quad \nu = 0, 1, 2, 3, \text{etc.}, \tag{67}$$

where x_ν are the radii of the dark fringes and

$$x_\mu = \sqrt{\mu R \lambda/2} ; \quad \mu = 1, 3, 5, \text{etc.}$$

where the x_μ are the radii of the bright fringes. The radius R of the surface can be computed from the measured values of the radii x_ν or x_μ.

16.13.1.3 The adoption of the theory of Fizeau fringes to Newton's fringes is, in itself, an approximation. The method of the sagitta should be regarded merely as a first approximation to the interpretation of Newton's fringes with extended sources of light. More critical investigations reveal that the choice of observation plane matters, as does also the location of the eye with respect to the points x_ν or x_μ.

16.13.1.4 In viewing both Fizeau and Newton's fringes, the tendency and practice is to focus upon the thin film between the interferometer surfaces S_1 and S_2.

16.13.1.5 For increased accuracy in using the sagitta method for determining the radius R_1, it is preferable to choose as the reference surface S_2 a spherical surface of known radius R_2 that departs only slightly from R_1. The "effective sagitta" s, Figure 16.15, is now given by $s = s_1 - s_2$ in which $x^2 = 2R_1 s_1 - s_1^2 = 2R_2 s_2 - s_2^2$. By neglecting s_1^2 and s_2^2 in comparison with $2R_1 s_1$ and $2R_2 s_2$, respectively, one obtains

$$s = \frac{x^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{x^2}{2}\left(\frac{R_2 - R_1}{R_1 R_2}\right). \tag{68}$$

Thus, from Equations (68) and (64), dark fringes occur at radii x_ν for which

$$x_\nu = \sqrt{\left(\frac{R_1 R_2}{R_1 - R_2}\right)\nu\lambda}\,, \tag{69}$$

a result that reduces to Equation (67) when $R_2 = \infty$. If R_1 and R_2 are nearly alike, one may set $R_1 R_2 = R_2^2$. Within the validity of this approximation,

$$R_2 - R_1 = \left(\frac{R_2}{x_\nu}\right)^2 \nu\lambda. \tag{70}$$

16.13.1.6 Systematic error of interpretation of Newton's fringes due to inadequacies of the sagitta method can be avoided or minimized, as will now be shown, by replacing one of the end-mirrors of the Tyman Green interferometer by the spherical surface S_1 as illustrated in Figure 16.16.

16.13.1.7 We suppose that the end-mirror S_1 has a large radius R and seek to compute R from the radii of the circular fringes seen about point O when the eye lens and telescope focus the plane $z = 0$ upon the retina. We may suppose for simplicity that pinhole H and the center C of spherical surface S_1 fall upon the axis of the instrument. We take plane $z = 0$ through point O as the plane of reference. The plane wave reflected from S_2 appears to return to the observer as a plane wave along the direction OZ. The wave reflected from S_1 appears (apart from spherical aberration produced on reflection) as a spherical wave that expands from point F located at distance R/2 behind point O. We suppose that the distances x are small enough that spherical aberration on reflection can be ignored.

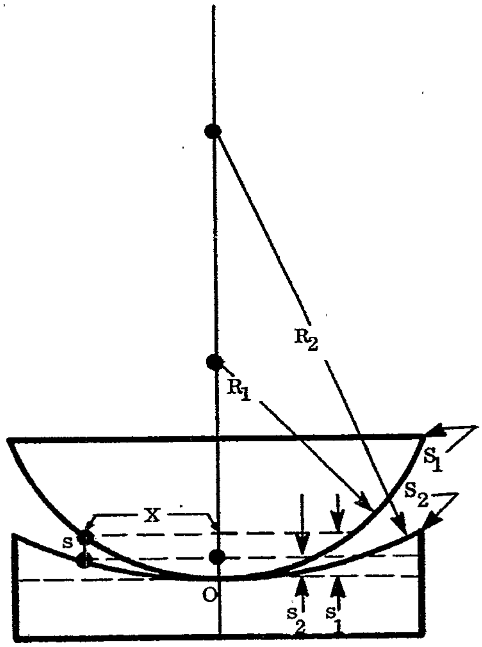


FIGURE 16.15--The Sagitta Method when the Reference Surface S_2 is a sphere.

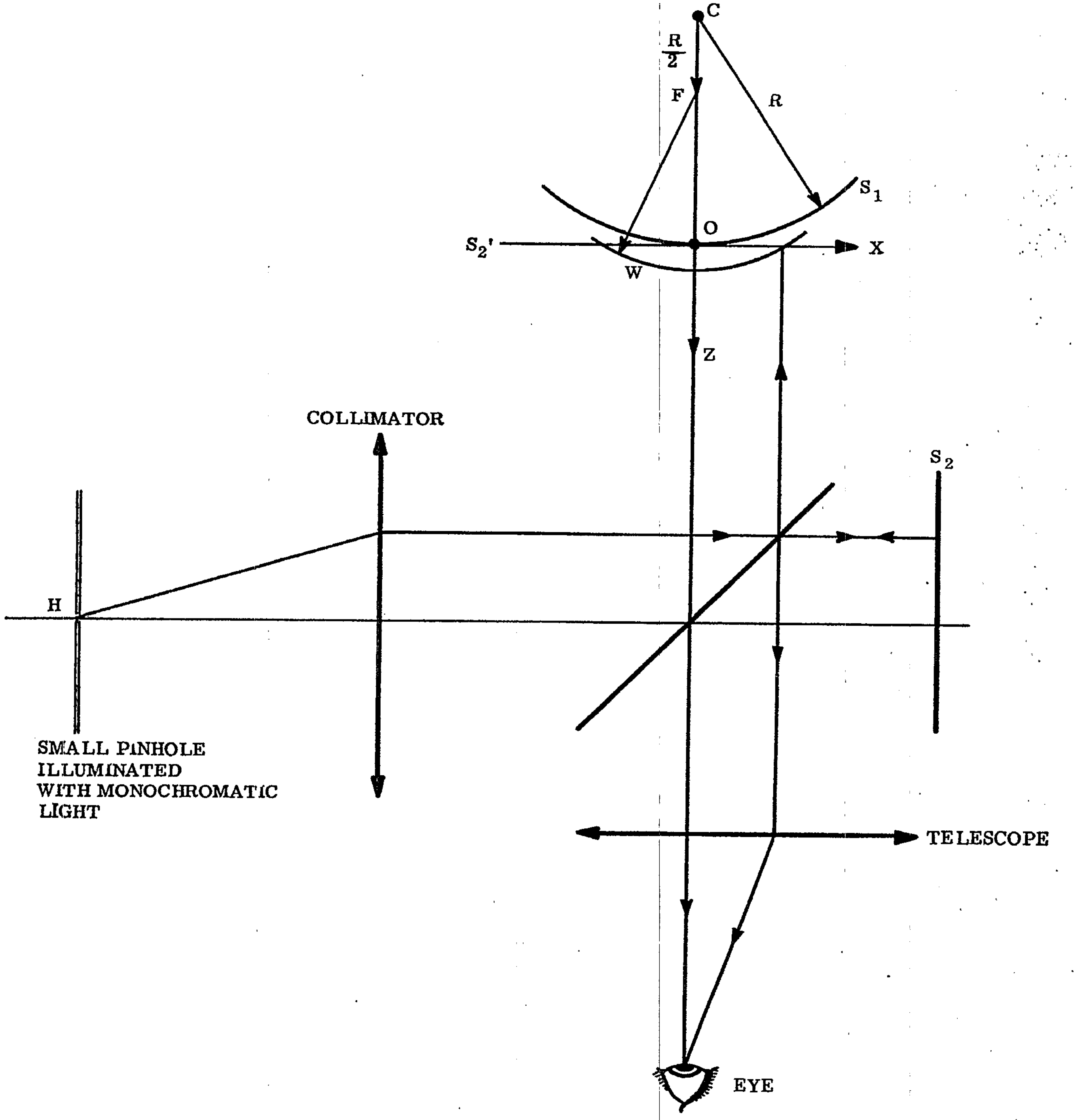


FIGURE 16. 16- Formation of Newton's Fringes with a Tyman Green Interferometer.

16.13.1.8 The plane wave returned to the observer has the form

$$E_2 = a_2 \cos(\delta_o + kz - \omega t) \tag{71}$$

where a_2 denotes the amplitude and δ_o has been added in order to account for the difference in phase change on reflection. The spherical wave returned to the observer has the form

$$E_1 = a_1 \cos\left\{\delta_1 + k\left[r^2 + \left(z - \frac{R}{2}\right)^2\right]^{1/2} - \omega t\right\} \tag{72}$$

in which

$$r \equiv (x^2 + y^2)^{1/2} . \tag{73}$$

On the circle $x^2 + y^2 = r^2$ in the plane of observation $z = 0$ the phase difference $\phi_1 - \phi_2$ between E_1 and E_2 is given by

$$\phi_1 - \phi_2 = \delta_1 - \delta_o + k\left(r^2 + \frac{R^2}{4}\right)^{1/2} . \tag{74}$$

But at point O, where $x = y = z = 0$, $\phi_1 - \phi_2$ must equal $-\delta_o$ because the separation of S_1 and S_2' is zero. Hence, with respect to the undetermined value of δ_1,

$$\delta_1 = -k\frac{R}{2} = -\pi\frac{R}{\lambda}$$

so that

$$\phi_1 - \phi_2 = -\delta_o - \frac{\pi R}{\lambda} + \frac{2\pi}{\lambda}\left[r^2 + \frac{R^2}{4}\right]^{1/2} . \tag{75}$$

The time-averaged energy density on circles of radii r the plane $z = 0$ is given by Equation (3) wherein $\phi_1 - \phi_2$ obeys Equation (75). It follows from Equations (3) and (75) that the fringes display maximum brightness at r-values for which

$$-\delta_o - \frac{\pi R}{\lambda} + \frac{2\pi}{\lambda}\left[r_\nu^2 + \frac{R^2}{4}\right]^{1/2} = \nu 2\pi ; \quad \nu = 0, 1, 2, 3, \text{etc.}, \tag{76}$$

and minimum brightness at r-values for which

$$-\delta_o - \frac{\pi R}{\lambda} + \frac{2\pi}{\lambda}\left[r_\mu^2 + \frac{R^2}{4}\right]^{1/2} = \mu\pi ; \quad \mu = 1, 3, 5, \text{etc.} \tag{77}$$

Equations (76) and (77) enable one to compute both δ_o and R from measured values of r_ν and r_μ in cases where δ_o is not known.

16.13.1.9 Either the Twyman Green interferometer or the Fizeau interferoscope of Figure 16. 2 may be used. With Fizeau interferoscopes, $\delta_o = \pi$ in Equations (76) and (77). With Twyman-Green interferometers, $\delta_o = 0$ when the end-mirrors are unsilvered or equally silvered surfaces of glass.

16.13.1.10 The exact form of Equations (76) and (77) will rarely be required. The following excellent approximation leads to a much simpler pair of working relations. We write

$$\left[r_\nu^2 + \frac{R^2}{4}\right]^{1/2} = \frac{R}{2}\left[1 + \frac{4r_\nu^2}{R^2}\right]^{\frac{1}{2}}$$

It will be impractical to utilize either the Twyman-Green or Fizeau interferometers unless the radius R of the test surface is so great that $1 \gg 4r_\nu^2/R^2$ and that

$$\frac{R}{2}\left[1 + \frac{4r_\nu^2}{R^2}\right]^{\frac{1}{2}} \doteq \frac{R}{2}\left[1 + \frac{2r_\nu^2}{R^2}\right] = \frac{R}{2} + \frac{r_\nu^2}{R} . \tag{78}$$

By combining Equations (76) and (77) with Equation (78), one obtains the simplified results

$$-\delta_o + \frac{2\pi r_\nu^2}{\lambda R} = \nu 2\pi ; \quad \text{(bright fringes)} \tag{79}$$

$$-\delta_o + \frac{2\pi r_\mu^2}{\lambda R} = \mu\pi ; \quad \text{(dark fringes)}. \tag{80}$$

If, for example, $\delta = \pi$ as in the Fizeau interferoscope,

$$2\pi r_\mu^2/\lambda R = (\mu + 1)\pi \quad \text{so that} \quad r_\mu^2 = \lambda R(\mu + 1)/2.$$

Consequently, for circular dark fringes

$$r_\mu = \sqrt{\lambda R(\mu + 1)/2}\,. \tag{81}$$

Since μ is an odd integer, $(\mu + 1)$ is an even integer, and $(\mu + 1)/2$ generates the integers 0, 1, 2, 3, etc. (To obtain the zero-value, one takes $\mu = -1$.)

16.13.1.11 Comparision of Equations (81) and (67) shows that they agree. This means that the sagitta method is more reliable as applied to measuring R in the Twyman-Green or the Fizeau interferometers than it is likely to be as applied to methods based upon Fizeau fringes or Newton's fringes. This conclusion is not surprising because the Fizeau and Twyman-Green interferometers utilize small sources of light and are constructed so that the observer is forced to view the fringes under conditions of normal incidence.

16.14 COMPLEX NUMBERS

16.14.1 Introduction.

16.14.1.1 Many of the following discussions are both shorter and more readily understood by employing complex numbers instead of the trigonometric functions. Only the most elementary properties of complex numbers will be needed.

16.14.1.2 One well known method of expressing a complex number Z is illustrated by the equation

$$Z = a + i\,b \tag{82}$$

wherein a and b are real numbers and $i = \sqrt{-1}$. The real numbers a and b are often called the real and imaginary parts, respectively. The so-called complex conjugate $\overline{Z}$ of Z is defined by the relation

$$\overline{Z} = a - i\,b. \tag{83}$$

It follows at once that

$$a = \frac{Z + \overline{Z}}{2} \equiv R_e(Z), \quad \text{the real part of } Z \tag{84}$$

and that

$$b = \frac{Z - \overline{Z}}{2i} \equiv I_m(Z), \quad \text{the imaginary part of } Z. \tag{85}$$

Furthermore,

$$|Z|^2 = Z\overline{Z} = a^2 + b^2 \; ; \quad (i^2 = -1) \tag{86}$$

where $|Z|$ is the absolute value or amplitude of Z, i. e., the length of Z as illustrated in Figure 16. 17

16.14.1.3 For our purposes the exponential form of Z is much to be preferred. Thus,

$$Z = |Z|\,e^{i\theta} \tag{87}$$

16.14.1.4 By definition,

$$|Z|\,e^{i\theta} = |Z|\,(\cos\theta + i\sin\theta) \tag{88}$$

where the angle θ, illustrated in Figure 16. 17, is called the argument of Z and written arg(Z). It follows by comparison of Equations (82) and (83) that

$$a = |Z|\cos\theta\,; \quad b = |Z|\sin\theta\,; \tag{89}$$

consequently,

$$\tan\theta = \frac{b}{a}\,. \tag{90}$$

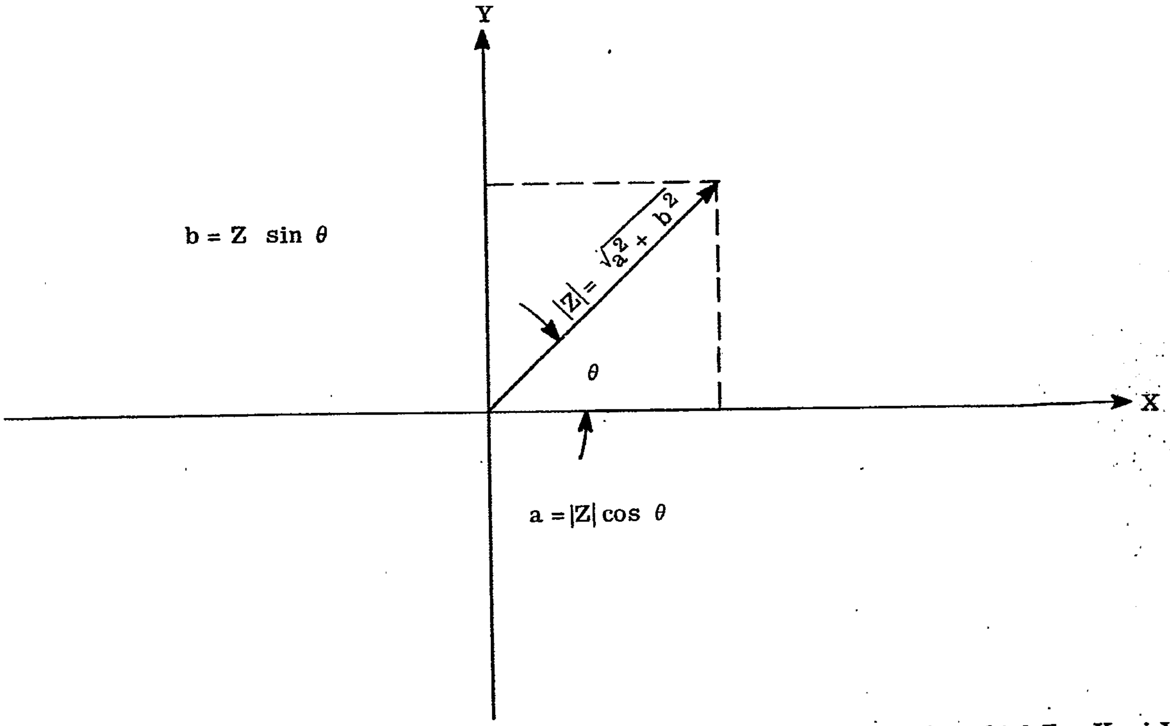


FIGURE 16. 17-Representation of complex numbers in the complex Z - plane for which Z = X + i Y.

16.14.1.5 That Equation (88) is a reasonable definition can be seen from the following considerations. From the series

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots$$

and

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

we obtain

$$\cos \theta + i \sin \theta = 1 + i\theta + \frac{i^2\theta^2}{2!} + \frac{i^3\theta^3}{3!} + \ldots$$

which is in the form of

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

wherein $x = i\theta$, and $e^x = e^{i\theta}$.

16.14.1.6 Given two complex numbers Z_1 and Z_2 in exponential form, their product Z is given by

$$Z = |Z_1| e^{i\theta_1} |Z_2| e^{i\theta_2} = |Z_1| |Z_2| e^{i(\theta_1 + \theta_2)} \tag{91}$$

The rule for multiplying two complex numbers is to multiply their amplitudes and to add their arguments. Similarly with respect to division,

$$Z = |Z_1| e^{i\theta_1} \Big/ |Z_2| e^{i\theta_2} = \frac{|Z_1|}{|Z_2|} e^{i(\theta_1 - \theta_2)} \tag{92}$$

16.14.1.7 Finally, if $Z = |Z| e^{i\theta}$,

$$\bar{Z} = |Z| e^{-i\theta}. \tag{93}$$

Consider, for example, the statement

$$E = ae^{i(\phi - \omega t)} = a \cos (\phi - \omega t) + i \sin (\phi - \omega t). \tag{94}$$

We see that the wave form of Equation (1) is the real part of E as expressed by the complex form of Equation (94). This means that, when desired, the instantaneous value of E can be computed as the real part, $R(E)$, of E as given by Equation (94). However, one's chief interests center finally upon the time-averaged energy density. From Equation (94)

$$|E|^2 = E\bar{E} = a^2. \tag{95}$$

By comparing Equations (95) and (2), we find that

$$2W = |E|^2 = E\bar{E} \tag{96}$$

where W is the time-averaged density. This property of the complex wave form is of great convenience.

16.14.1.8 Suppose that the complex wave traverses a medium whose amplitude transmittance is τ and whose phase transmittance (optical path) is nd. We can write the transmittance of this medium in the complex form

$$T = \tau e^{ind} \tag{97}$$

If E is given by Equation (94) upon entry into the medium, then if E' denotes the value of E as the wave leaves the medium

$$E' = TE = \tau a e^{i(\phi + nd - \omega t)} \tag{98}$$

Similarly, if the wave corresponding to Equation (94) is reflected from an interface between two media

$$E' = \rho E = |\rho| e^{i(\phi + \psi - \omega t)} \tag{99}$$

in which $\rho = |\rho| e^{i\psi}$ wherein ρ denotes amplitude reflectance and ψ denotes the phase change upon reflection.

16.15 TRANSMITTANCE OF PLANE PARALLEL PLATES

16.15.1 Introduction.

16.15.1.1 The simplified treatment of paragraph 16.9 applies with excellent approximation to plates whose surfaces have low reflectance. As the reflectance of the surfaces increases, the effects of the inter-reflected beams ultimately dominate and exert, as we shall see, profound effects upon the distribution of energy density in the observed fringes. The most conspicuous of these effects is a pronounced sharpening of the fringes to the point where they can appear as narrow bright lines on a dark background in transmitted light. These narrow fringes can be utilized to obtain more accurate measurements of surface irregularities, etc., than is possible with the sinusoidal fringes that are produced by double beam interferometers, such as the Michelson interferometer or the Fizeau interferoscope.

16.15.1.2 The theory of this paragraph applies directly to the Fabry-Perot and related multiple beam interferometers.

16.15.1.3 With respect to Figure 16.18:

n_1 = refractive index of the plate
n_o = refractive index of medium of incidence
n_2 = refractive index of last medium
d = thickness of the plate
i = angle of incidence
i' = angle of refraction
τ_1 = internal amplitude transmittance of the plate
$\tau_{0,1}$ = amplitude transmittance from the 0th into the 1st medium
$r_{1,0}$ = amplitude reflectance from the 1st upon the 0th medium
$\delta_{1,0}$ = phase change on reflection associated with $r_{1,0}$
$\tau_{1,2}$ = amplitude transmittance from 1st into the 2nd medium
$r_{1,2}$ = amplitude reflectance from the 1st upon the 2nd medium
$\delta_{1,2}$ = phase change on reflection associated with $r_{1,2}$

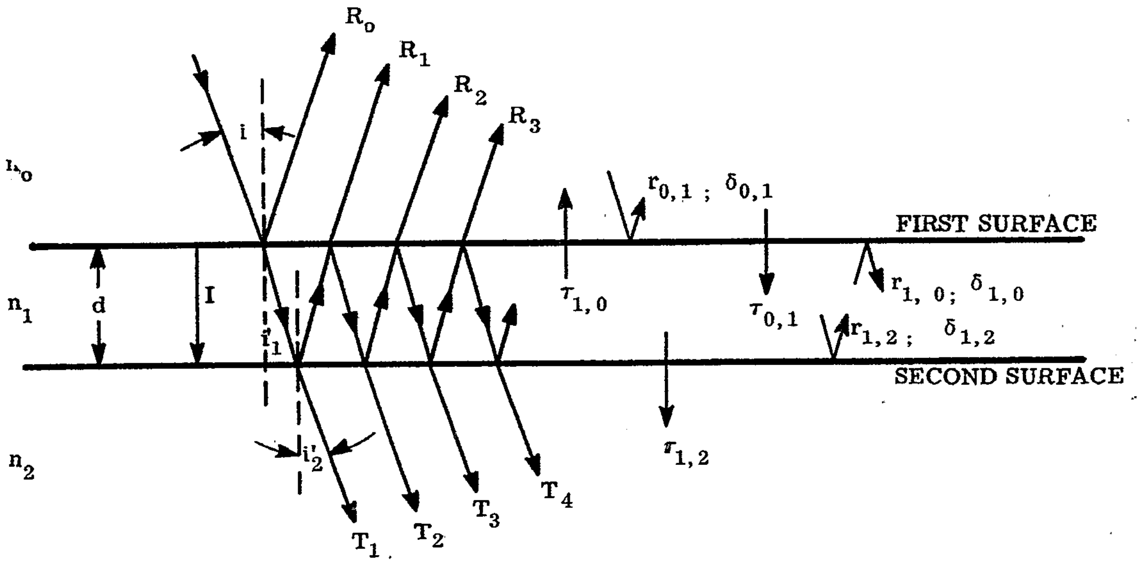


FIGURE 16. 18-Convention with respect to the transmitted beam in a plate or Fabry-Perot interferometer.

16.15.1.4 We bear in mind that $\tau_{0,1}$, $\rho_{1,0} = r_{1,0}\, e^{i\delta_{1,0}}$, $\tau_{1,2}$ and $\rho_{1,2} = r_{1,2}\, e^{i\delta_{1,2}}$ are Fresnel coefficients that depend in general upon i and upon whether the incident E-vector vibrates in, or perpendicular to the plane of the paper.

$$n_o \sin i = n_1 \sin i'_1 = n_2 \sin i'_2 \,. \tag{100}$$

The optical path difference between any two rays T_j and T_{j+1} will be $2n_1 d \cos i'_1$ (see paragraph 5.10). Let

$$\alpha \equiv \frac{4\pi n_1}{\lambda} d \cos i'_1 + \delta_{1,0} + \delta_{1,2} \tag{101}$$

and let β be the optical path for the directly transmitted beam T_1. Then, under the supposition that the incident beam has the amplitude unity,

$$T_1 = \tau_{0,1}\ \tau_1\ \tau_{1,2}\, e^{i\beta}$$

$$T_2 = \tau_{0,1}\ \tau_1^3\ \tau_{1,2}\, e^{i\beta}\ r_{1,2}\ r_{1,0}\ e^{i\alpha}$$

$$T_3 = \tau_{0,1}\ \tau_1^5\ \tau_{1,2}\, e^{i\beta}\ (r_{1,2}\ r_{1,0})^2\ e^{i\,2\alpha}\ ;\ \text{etc.}$$

If we consider N inter-reflections so that there are N emergent rays T_j, the emergent wave is now determined from the scalar quantity

$$E = e^{-i\omega t} \sum_{\nu=1}^{N} T_\nu = \tau_{0,1}\ \tau_1,\ \tau_{1,2}\ e^{i\beta} e^{-i\omega t} \sum_{\nu=0}^{N} A^\nu\ e^{i\nu\alpha} \tag{102}$$

where

$$A = \tau_1^2\ r_{1,0},\ r_{1,2} \leq 1. \tag{103}$$

But

$$\sum_{\nu=0}^{N} A^\nu\ e^{i\nu\alpha} = \frac{1 - A^{N+1}\ e^{i\alpha(N+1)}}{1 - A\ e^{i\alpha}} \tag{104}$$

Therefore,

$$E = \tau_{0,1},\ \tau_1,\ \tau_{1,2}\ e^{i(\beta - \omega t)} \frac{1 - A^{N+1}\ e^{i\alpha(N+1)}}{1 - A\ e^{i\alpha}}. \tag{105}$$

16.15.1.5 The time-averaged energy density 2W is proportional to $|E|^2 = E\ \bar{E}$. It is obtained in a straightforward manner from Equation (105). The result is

$$2W = (\tau_{0,1},\ \tau_1,\ \tau_{1,2})^2 \left\{ \frac{1 - 2\ A^{N+1}\ \cos\ [(N+1)]\alpha\ + A^{2(N+1)}}{1 - 2\ A\ \cos\ \alpha\ +\ A^2} \right\}. \tag{106}$$

wherein α and A are given by Equations (101) and (103), respectively.

16.15.1.6 In a thick plate the number N of inter-reflections is restricted by the length of the incident wave train or by the tendency of each successive reflection to "walk" the beam out through the ends of the plate. However, with thin films, such as soap films, or with evaporated films one is usually justified in setting $N = \infty$. Whenever one can accept the approximation $N = \infty$, the time-averaged energy density W in the transmitted beam is given by the simpler expression

$$2W = \frac{(\tau_{0,1}\ \tau_1,\ \tau_{1,2})^2}{1 - 2\ A\ \cos\ \alpha + A^2}. \tag{107}$$

16.15.1.7 With respect to both Equations (106) and (107), major maxima occur in the transmitted fringes when

$$\alpha = \nu\ 2\ \pi;\quad \nu = 0,\ 1,\ 2,\ 3,\ \text{etc.} \tag{108}$$

This result can be expected intuitively; for it requires that all rays T_j of Figure 16.18 shall emerge in phase.

16.15.1.8 The integers ν are often called spectral orders.

16.15.1.9 Equation (106) for N transmitted rays T_j differs from Equation (107) in that it predicts the existence of $N + 1$ subsidiary maxima between any two consecutive spectral orders ν and $\nu + 1$.

16.15.1.10 When $A = \tau_1^2\ r_{1,0}\ r_{1,2}$ becomes small in Equations (106) and (107),

$$2W \rightarrow \frac{(\tau_{0,1}\ \tau_1\ \tau_{1,2})^2}{1 - 2\ A\ \cos\ \alpha} \rightarrow (\tau_{0,1}\ \tau_1\ \tau_{1,2})^2\ (1 + 2\ A\ \cos\ \alpha). \tag{109}$$

This means that the transmitted fringes assume the sinusoidal distributions typical of double beam interferometers when A becomes small due to reduction of the internal transmittance τ_1 of the plate or of the amplitude reflectances $r_{1,0}$ and $r_{1,2}$ of its surfaces. Contrast in the transmitted fringes will be poor when A is so small that Equation (109) is an acceptable approximation to Equation (107).

16.15.1.11 It is more difficult to demonstrate that Equations (107) and (106) predict the appearance of sharp fringes as A approaches unity. Let the energy density W be plotted against α as in Figure 16.19. At $\alpha = \nu 2 \pi$,

$$W = W_{max} = \frac{B^2}{(1 - A)^2} \tag{110}$$

in which $B = \tau_{0,1} \tau_1 \tau_{1,2}$. We wish to find the neighboring value of α for which $W = W_{max}/2$. Set

$$\alpha = \nu 2 \pi + \Delta \alpha \tag{111}$$

and suppose that $\Delta\alpha$ is so small that $\cos \alpha = \cos \nu 2 \pi \cos \Delta\alpha = 1 - (\Delta\alpha)^2/2$. Then for $W = W_{max}/2$ from Equation (107), $B^2/\left[1 - 2A + A^2 + A (\Delta\alpha)^2\right] = B^2/\left[2 (1 - A)^2\right]$ so that $A (\Delta\alpha)^2 = (1 - A)^2$. Hence,

$$\Delta \alpha = \frac{1 - A}{\sqrt{A}} \tag{112}$$

where $\Delta \alpha$ is the increment that must be added to $\alpha = \nu 2 \pi$ in order to drop W from W_{max} to $W_{max}/2$. If α is increased by 2π, the next fringe for which $W = W_{max}$ is obtained. In other words, the fringe-width is 2π in terms of α. We define

$$w = \frac{\Delta \alpha}{2 \pi} = \frac{1}{2 \pi} \frac{1 - A}{\sqrt{A}} \tag{113}$$

and call it the optical half-width of the Fabry-Perot fringes. We see that this optical half-width decreases rapidly as A approaches unity. If, for example, $A = 0.9$, $2w = 0.032$. This means that the width 2w, Figure 16.19, is approximately 0.03 times the width from one bright fringe to the next. The fringes become exceedingly sharp as A approaches unity. A-values of 0.9 are obtained easily by silvering the two surfaces

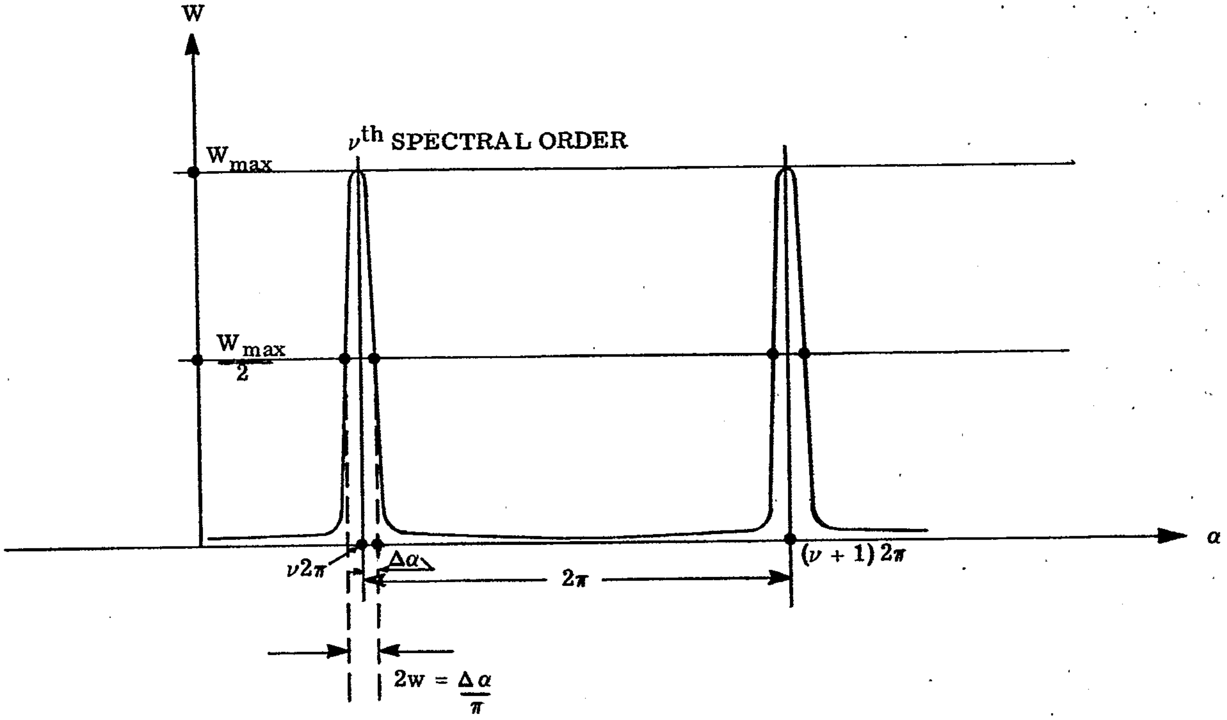


FIGURE 16.19- The Sharpness Quality of Multiple Beam Interference Fringes.

of the plate.

16.15.1.12 The limiting sharpness of the multiple beam fringes depends ultimately upon freedom from absorption. As a high reflecting coating, silver has remarkably low absorption. It is not difficult to obtain evaporated films of silver that have absorptions less than 5 per cent even when the film is practically opaque. Whereas much lower absorptions are possible with silver, the use of high reflecting, multi-layered films is becoming more common when the narrowest half-widths are required.

16.15.1.13 Two methods for viewing the multiple beam interference fringes that are transmitted by a plate are illustrated in Figures 16. 20 and 16. 21. Sharp, circular fringes will be seen provided that the surfaces of the plate are sufficiently parallel and silvered. Since n_1, d, and λ are fixed, it follows at once from Equation (101) that the sharp bright fringes are fringes of equal inclination, i. e., the angle of refraction i_1' is constant along each fringe. When the thickness d of the plate or film is large, the number of circular fringes becomes so great that the determination of their spectral order ν is difficult.

16.16 REFLECTANCE FROM PLANE PARALLEL PLATES

16.16.1 Introduction.

16.16.1.1 The dark fringes usually appear sharp in the reflected family. However, it is not necessarily true that a dark fringe must appear in the reflected family of fringes at values of α for which a bright fringe occurs in the transmitted family.

16.16.1.2 With respect to Figure 16.18, let $r_{0,1}$ and $\delta_{0,1}$ denote amplitude reflectance and phase change on reflection for a beam incident from the 0th medium. Then,

$$\begin{aligned} R_0 &= r_{0,1}\, e^{i\delta_{0,1}} \\ R_1 &= \tau_{0,1}\, \tau_{1,0}\, \tau_1^2\, r_{1,2}\, e^{i(\alpha - \delta_{1,0})} \\ R_2 &= \tau_{0,1}\, \tau_{1,0}\, \tau_1^4\, r_{1,2}^2\, r_{1,0}\, e^{i(2\alpha - \delta_{1,0})} \\ R_3 &= \tau_{0,1}\, \tau_{1,0}\, \tau_1^6\, r_{1,2}^3\, r_{1,0}^2 \cdot e^{i(3\alpha - \delta_{1,0})}, \text{ etc.} \end{aligned} \tag{114}$$

Therefore,

$$R = \sum_{\nu=0}^{N} R^{\nu} = r_{0,1}\, e^{i\delta_{0,1}} + C\, e^{(\alpha - \delta_{1,0})} \sum_{\nu=0}^{N-1} A^{\nu}\, e^{i\nu\alpha} \tag{115}$$

in which α and A are defined by Equations (101) and (103), R is a complex number that determines the amplitude and phase of the reflected beam and

$$C \equiv \tau_{0,1}\, \tau_{1,0}\, \tau_1^2\, r_{1,2}. \tag{116}$$

Comparison of Equations (102) and (115) shows that Equation (115) contains the additional term $r_{0,1}\, e^{i\delta_{0,1}}$ due to the first reflection R_0 of Figure 16. 18. It is the presence of this extra term that complicates the nature and the interpretation of the reflected fringes.

16.16.1.3 Suppose that α has any one of the values $\nu 2\pi$ of Equation (108), the condition for bright fringes in the transmitted beam. Then from Equation (115)

$$\begin{aligned} R &= r_{0,1}\, e^{i\delta_{0,1}} + C\, e^{-i\delta_{1,0}} \sum_{\nu=0}^{N-1} A^{\nu} \\ &= r_{0,1}\, e^{i\delta_{0,1}} \quad 1 + C\, e^{-i(\delta_{1,0} + \delta_{0,1})} \sum_{\nu=0}^{N-1} A^{\nu} \end{aligned} \tag{117}$$

We see that $|R|$ will be minimum when $\alpha = \nu\, 2\pi$, provided that with respect to the phase changes $\delta_{0,1}$ and $\delta_{1,0}$ on reflection at the first surface of the plate

$$\delta_{1,0} + \delta_{0,1} = \mu\pi \tag{118}$$

where μ is an odd integer. In other words, dark reflected fringes will occur at the same α - values as bright transmitted fringes, provided that the sum of the phase changes on reflection for incidence from opposite directions upon the first surface, Figure 16. 18, is an odd number of half-wavelengths. This condition is rarely fulfilled. Consequently, one has to expect that the reflected fringes will be darkest at α - values that differ suitably from $\alpha = \nu\, 2\pi$ where ν is an integer. However, this complication does not detract from the utility of

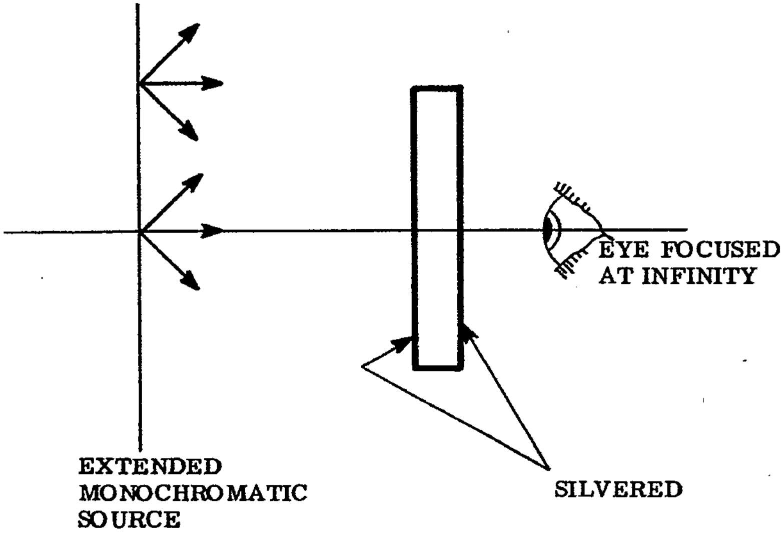


FIGURE 16. 20-Simple Parallel Plate Interferometer.

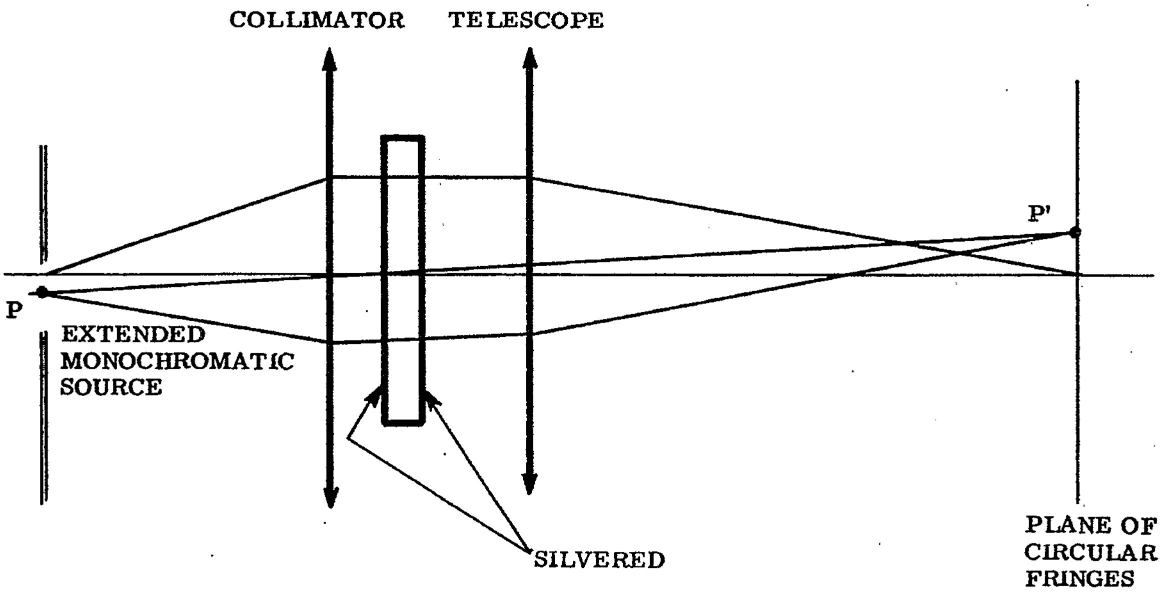


FIGURE 16. 21- The Fabry-Perot Interferometer.

the reflected fringes, except in those cases in which it leads to fringes that are only slightly darker than the background.

16.16.1.4 Sharp reflected fringes can be observed, for example, by replacing the elements bearing surfaces S_1 and S_2 of Figure 16.11 by a plane parallel plate whose major surfaces are suitably silvered, aluminized, etc. The eye is preferably focused for infinity.

16.17 MULTIPLE BEAM INTERFERENCE FRINGES FROM SLIGHTLY INCLINED SURFACES

16.17.1 General.

16.17.1.1 Let a wavefront V be incident upon the wedge formed between two reflecting surfaces that have the small included angle α as illustrated in Figure 16.22. Wavefronts V_0, V_1, V_2, etc., inclined at the angles 0, 2α, 4α, etc., will emerge from the wedge after an appropriate number of inter-reflections within the wedge. The corresponding emergent rays are indicated by T_o, T_1, T_2, etc. A series of coherent, plane waves are formed in this manner by inter-reflections within the wedge.

16.17.1.2 Let

$t_1 \equiv$ amplitude transmittance of surface S_1;
$t_2 \equiv$ amplitude transmittance of surface S_2;
$r_1 \equiv$ amplitude reflectance of surface S_1;
$r_2 \equiv$ amplitude reflectance of surface S_2;
$\delta_1 \equiv$ phase change on reflection at surface S_1;
$\delta_2 \equiv$ phase change on reflection at surface S_2.

16.17.1.3 We choose the X-axis along OP and the Z-axis parallel to PT_o and suppose that the amplitude of the incident wavefront is unity. We note that such phase changes as may occur upon transmission through surfaces S_1 and S_2 can be ignored since they alter all of the emergent waves equally. The space between S_1 and S_2 is assumed to be nonabsorbing.

16.17.1.4 The emergent wave propagated along PT_o, i.e., along Z, has the complex form

$$T_o = t_1 t_2 e^{iknz} e^{-i\omega t}$$

The wave emergent along PT_1 has the form

$$T_1 = t_1 t_2 r_1 r_2 e^{i(\delta_1+\delta_2)} e^{ikn[x \sin 2\alpha + z \cos 2\alpha]} e^{-i\omega t}$$

Similarly,

$$T_2 = t_1 t_2 (r_1 r_2)^2 e^{i2(\delta_1+\delta_2)} e^{ikn[x \sin 4\alpha + z \cos 4\alpha]} e^{-i\omega t};$$

$$T_3 = t_1 t_2 (r_1 r_2)^3 e^{i3(\delta_1+\delta_2)} e^{ikn[x \sin 6\alpha + z \cos 6\alpha]} e^{-i\omega t};$$

etc.,

16.17.1.5 Introduce

$$R \equiv r_1 r_2;$$

$$\tau \equiv t_1 t_2;$$

$$\phi \equiv \delta_1 + \delta_2. \tag{119}$$

Then

$$T = \sum_{\nu=0}^{N} T^{\nu} = \tau e^{-i\omega t} \sum_{\nu=0}^{N} R^{\nu} e^{i\nu\phi} e^{ikn[x \sin(2\nu\alpha) + z \cos(2\nu\alpha)]} \tag{120}$$

where T specifies the amplitude and phase determined by the interference of the emergent waves T_o, T_1, ... , T_N. The fringes described by Equation (120) are of a type far more general than those ordinarily used. We obtain the conventional type multiple beam fringes formed by a wedge by supposing that the angle α of the wedge is so small that

$$\sin(2\nu\alpha) \to 2\nu\alpha; \quad 0 \leqq \nu \leqq N. \tag{121}$$

$$\cos(2\nu\alpha) \to 1; \quad 0 \leqq \nu \leqq N.$$

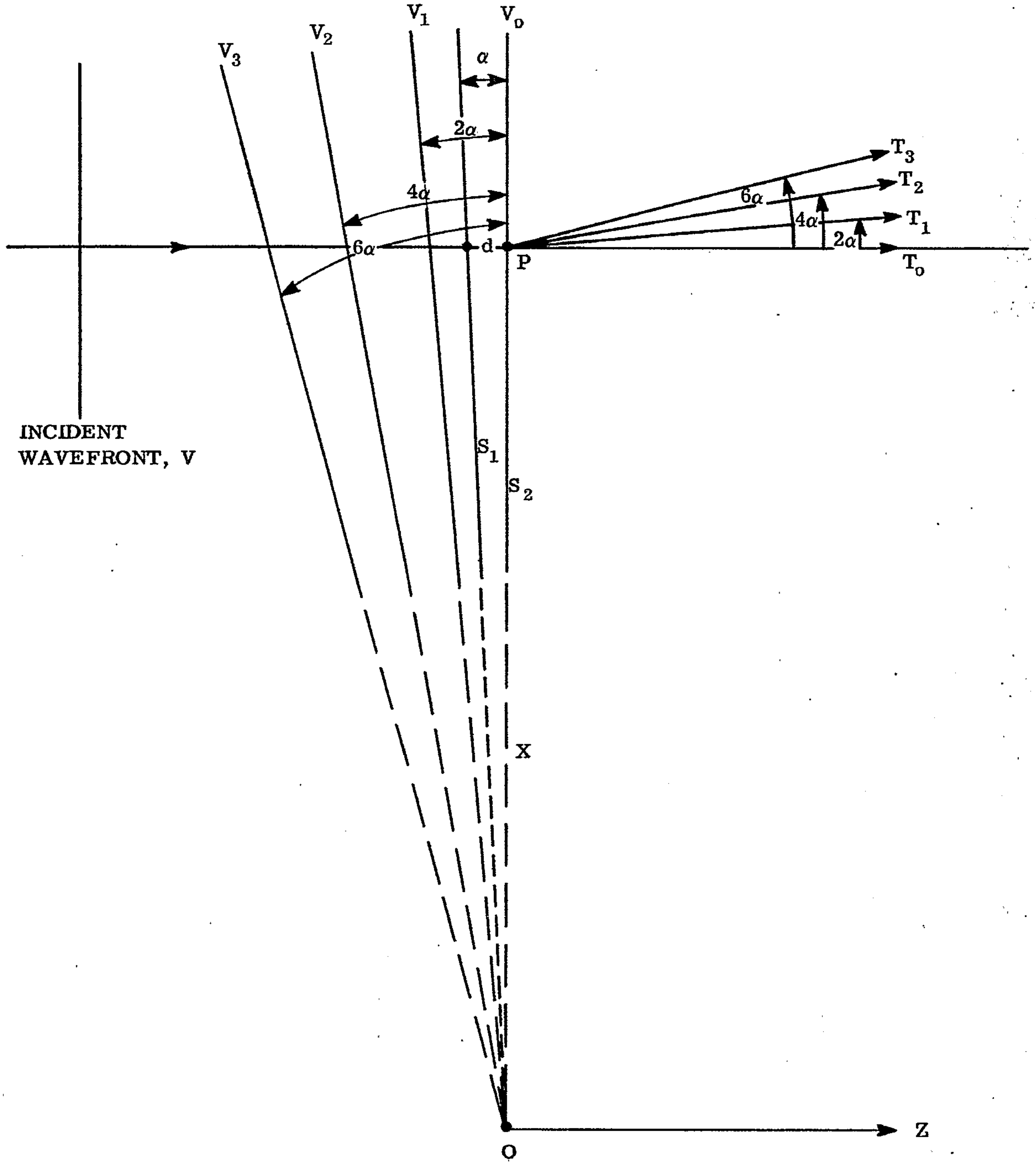


FIGURE 16. 22- Multiple Reflections in two reflecting surfaces S_1 and S_2 .

From Equations (120) and (121)

$$T = \tau \, e^{-i\omega t} \, e^{iknz} \sum_{\nu = 0}^{N} R^{\nu} \, e^{i\nu\,(\phi + 2knx\alpha)} \tag{122}$$

Applying Equation (104) to Equation (122), we obtain

$$T = \tau \, e^{-i\omega t} \, e^{iknz} \, \frac{1 - R^{N+1} \, e^{i\,(N+1)\,(\phi + 2knx\alpha)}}{1 - R \, e^{i\,(\phi + 2knx\alpha)}} \tag{123}$$

16.17.1.6 The time-averaged energy density W_T in the fringes seen on transmission is given by

$$2W_T = |T|^2 = \tau^2 \, \frac{1 - 2R^{N+1} \cos\,[(N+1)(\phi + 2\,knx\,\alpha)] + R^{2(N+1)}}{1 - 2R \cos\,(\phi + 2\,knx\,\alpha) + R^2} \tag{124}$$

in which

$$\tau \equiv t_1 t_2 ; \quad R \equiv r_1 r_2 ; \quad \phi \equiv \delta_1 + \delta_2 ; \tag{119}$$

$k \equiv 2\pi/\lambda$; and n is the refractive index of the medium within the wedge. ϕ is the sum of the phase changes on reflection at the surfaces S_1 and S_2 of the wedge. α is the angle of the wedge. The result of Equation (124) is independent of z (which suggests most strongly that the fringes are not necessarily localized within the wedge). However, it should be remembered that the requirement of Equation (121) is unlikely to be met in actual practice when the included number of inter-reflections N is high. Dependence of the fringe system upon the plane z of observation must be expected from Equation (120) when one is not entitled to set $\cos\,(2\,\nu\,\theta) = 1$.

16.17.1.7 A common method for obtaining and viewing transmitted multiple beam fringes in a wedge is illustrated in Figure 16. 23. The rays PT_0, PT_1, PT_2, etc., of Figure 16. 22 form images H_0, H_1, H_2, H_3, etc., of the pinhole H at the second focal plane of the objective. The number N of inter-reflections is frequently restricted by the diaphragm stop D of the objective, i. e., by the numerical aperture of the objective. In Figure 16. 23 rays from the zero order ($\nu = 0$) pass through H_0 ; rays from the 1st order ($\nu = 1$) pass through

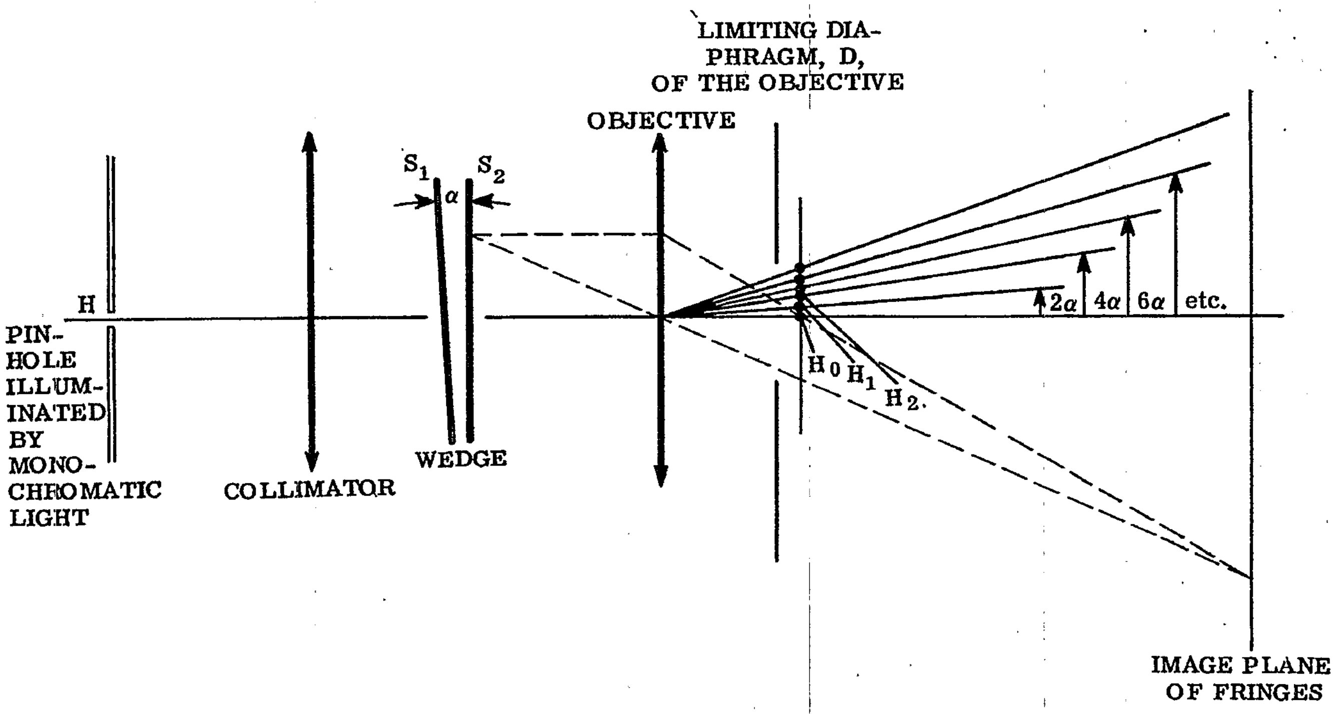

LIMITING DIAPHRAGM, D, OF THE OBJECTIVE
OBJECTIVE
S1
S2
α
H
PIN-HOLE ILLUMINATED BY MONOCHROMATIC LIGHT
COLLIMATOR
WEDGE
2α
4α
6α
etc.
H0
H1
H2
IMAGE PLANE OF FRINGES

FIGURE 16. 23- Method of producing and viewing Transmitted, Multiple Beam Fringes in a wedge.

H_1 , etc. Rays belonging to the seventh order ($\nu = 7$) are interrupted by the diaphragm. Thus, with respect to Equation (124), one would have $\nu_{max} = N = 6$. Apart from restricting the possible number of spectral orders N that get to the image plane, the objective may be regarded as a means of observing the object plane which is usually selected at, or within, the wedge where z approaches zero. The pinhole images H_ν, that correspond to the spectral orders ν, are easily seen by viewing the back of the objective, provided the system is in proper adjustment and that α is neither too small nor too large. The image plane is frequently viewed with an eyepiece. A microscope forms an excellent means for viewing the transmitted fringes. One has only to replace the conventional substage condensor by a more suitable lens to act as collimator. The selected pinhole H should be small enough so that it does not reduce the sharpness of the multiple beam fringes as determined experimentally.

16.17.1.8 We now return to complete our interpretation of Equation (124). Comparison of Equations (106) and (107) for multiple beam fringes with plane parallel plates shows that they are very similar to Equation (124). Most of the conclusions drawn in paragraph 16.15 apply again with minor modifications or qualifications. For example, we may conclude at once that bright fringes will occur when

$$\phi + \frac{4\pi}{\lambda} nx\alpha = \nu\, 2\pi . \qquad (125)$$

It will be seen from Figure 16.22 that

$$x\alpha = d \qquad (126)$$

where d is the thickness of the wedge at the point P under observation. Hence, we may rewrite Equation (125) in the well known form

$$2nd = \nu\lambda - \frac{\lambda\phi}{2\pi} \qquad (127)$$

in which ϕ, expressed in radians, is the sum of the phase changes on reflection at the surfaces S_1 and S_2 of the wedge. ϕ is in general a function of the wavelength. Again we observe that each fringe is the locus of points x for which the optical path nd is constant. The fringe width $|\Delta x| = h$ must be, according to Equation (125), that value of $|\Delta x|$ for which $4\pi n\alpha\, |\Delta x| / \lambda = 2\pi$. Therefore, the fringe width h is given by

$$h = |\Delta x| = \frac{\lambda}{2n\alpha} . \qquad (128)$$

Comparison of Equation (128) with Equation (12) shows that when the refractive indices n of the space between the reflecting surfaces are alike, the fringe widths are the same, whether one is using a Fizeau type interferometer or the multiple beam interferometer.

16.17.1.9 With respect to Figure 16.22, reflected plane waves emerge from the wedge and are propagated along the negative Z-direction. Corresponding to Equation (120), a series R for the reflected fringes is obtained. As in Equation (115) for parallel plates (case $\alpha = 0$), the series for R is complicated by the term R_0 that corresponds to direct reflection from the first surface of the wedge. In general, the remarks and conclusions of paragraph 16.16 also apply to the multiple beam fringes formed by reflection from a wedge for which $\alpha \neq 0$. The narrow reflected fringes are likely to be dark. A useful method for observing reflected multiple beam fringes is illustrated in Figure 16.24. The pinhole is placed at the first focal plane of the objective. The images H_0 , $H_1, \ldots, H_N$ of the pinhole H formed by the light belonging to the spectral orders ν fall along a straight line. When an undue amount of parasitic light is present at the plane of H_0 , $H_1, \ldots, H_N$, contrast in the fringes can be improved markedly by inserting at this plane a diaphragm with a slit which is oriented to pass the spectral orders. It is possible also to block the spectral order $\nu = 0$ by blocking the light in the image H_0 . When this is done, the reflected fringes have the appearance of the transmitted fringes — in fact, these narrow, bright, reflected fringes now obey Equation (125).

16.18 MEASUREMENTS WITH MONOCHROMATIC LIGHT

16.18.1 Introduction.

16.18.1.1 The effects of thin films upon the phase change introduced into a wave that traverses the optical system are being considered by some designers as an integral portion of the optical design of complex, high quality systems that contain many coated elements. The multiple beam interferometer is used frequently for measuring the thickness of thin films. The following principles belong to a method that has been applied to many different types of thickness measurements notably by S. Tolansky.

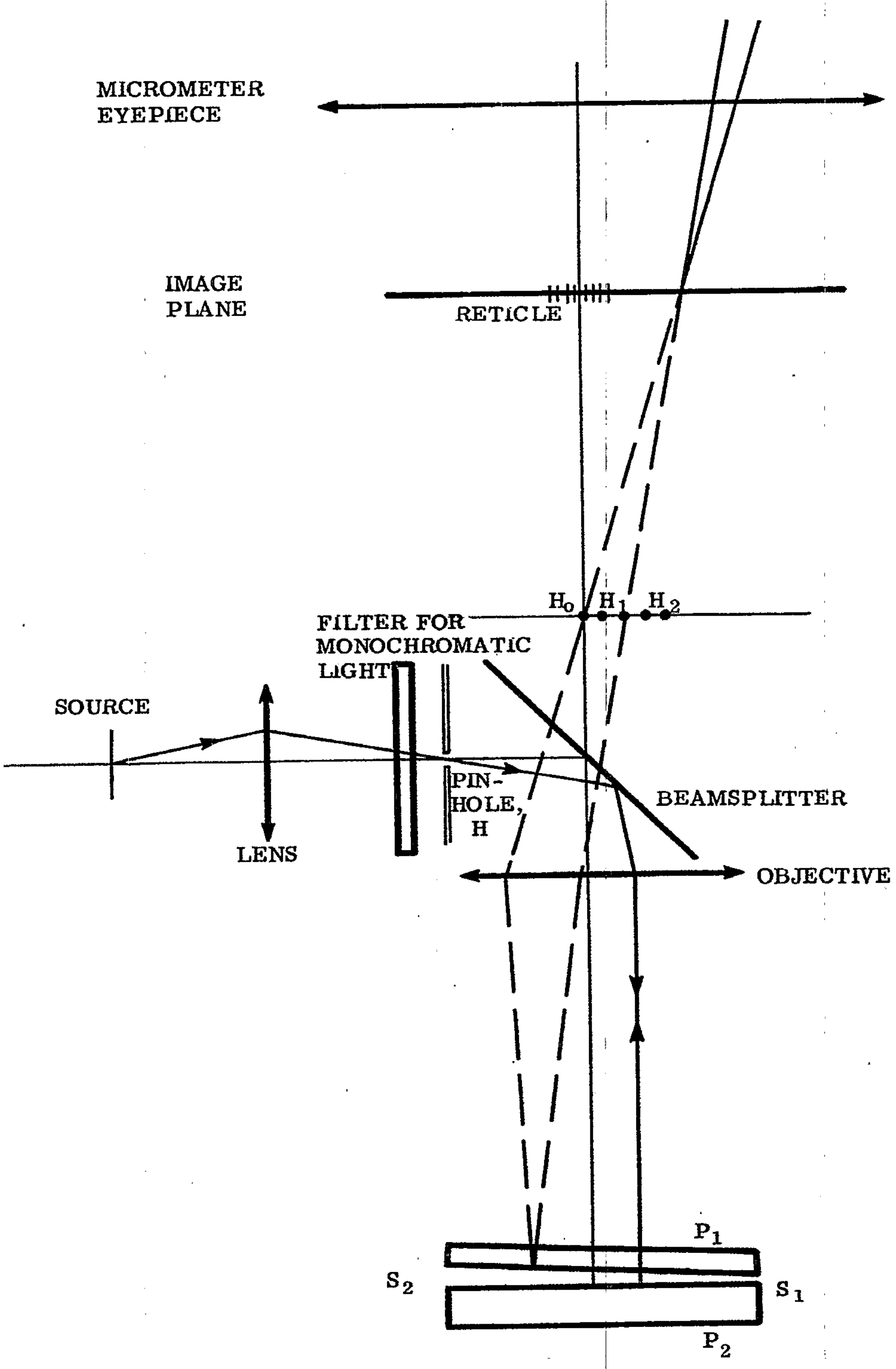


FIGURE 16. 24--Microscope for viewing reflected fringes under vertical illumination.

16.18.1.1 The preferred arrangement for measuring thicknesses of thin films utilizes multiple beam fringes that are formed by reflection as illustrated in Figure 16. 24. A micrometer eyepiece, containing any suitable reticule, is needed for measuring fringe widths and the fringe shifts that occur at the edge of a film that has been deposited upon surface S_2 and covered with a uniform coating of, say, silver as illustrated in Figure 16. 25. Evaporated coatings of silver and other metals produce a sharp step, whose height is equal to that of the film. The evaporated overcoating must be sufficiently opaque so that the phase changes on reflection at S_2 are not changed by the presence of the substrate or the film. The optically flat surface S_1 must be placed in close contact with surface S_2 in order to obtain reliable measurements of the thickness of the film. The usual practice is to lay plate P_1 directly upon plate P_2, Figure 16. 24, after making certain that no large dust particles are present to increase the separation between the silvered surfaces. It is good practice to make the fringes approximately perpendicular to the edge AB as in Figure 16. 26.

16.18.1.2 Let t denote the thickness of the film. We shall now show that

$$t = \frac{1}{2} \frac{\Delta x}{n h} \lambda \tag{129}$$

where Δx and h are respectively, the fringe shift and fringe width determined with the aid of the micrometer eyepiece (see Figure 16. 26). It is presumed that t is so small that the fringe shift is less than one fringe width. (This method is not well suited to measure thicknesses for which the fringe shifts Δx exceed the fringe width.) We have seen that a fringe is the locus of points Δx for which the separation d of the reflecting surfaces is constant. If then, a fringe is located at the point x in the absence of the film, it will move to a point $x + \Delta x$ on the film so as to keep d constant in the manner illustrated in Figure 16. 27. Since the angle α between S_1 and S_2 is to be small,

$$t = \alpha \Delta x . \tag{130}$$

But from Equation (128), $\alpha = \lambda/2 n h$. Substitution of this value of α into Equation (130) gives Equation (129) directly. The wedge between surfaces S_1 and S_2 is ordinarily air so that $n = 1$. This simple argument leading to Equation (129) applies to both the reflected and the transmitted fringes.

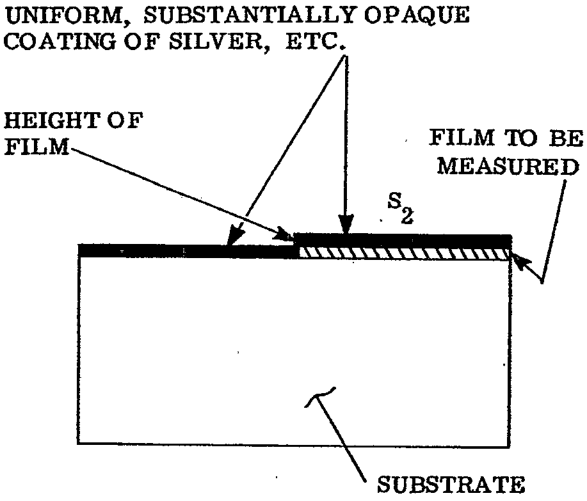


Figure 16. 25 - The usual method of preparing the sample film for thickness measurement in the Multiple Beam Interferometer.

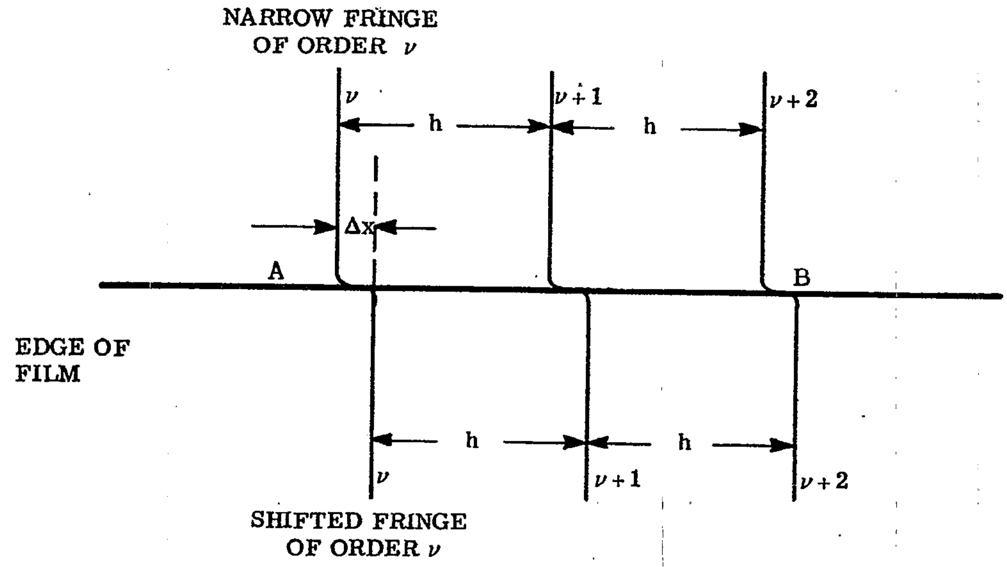


FIGURE 16. 26-Appearance of the narrow fringes when the thickness of the film is a small fraction of a wavelength and the film occupies the portion below the edge AB. If the surfaces S_1 and S_2 are optically flat, the indicated fringe widths, h, will be alike within the expiramental error of measuring h.

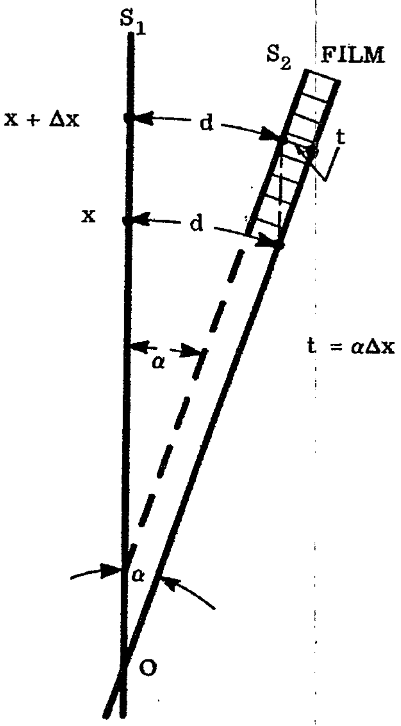


FIGURE 16. 27- Movement of an interference frings, x, to the position, $x + \Delta x$, by the introduction of a film at thickness, t.

16.18.1.3 One soon finds that the attainable precision is restricted by the roughness of the polished glass surfaces that ordinarily serve as the reflectors. These surfaces present, so to speak, a mountainous terrain whose peaks and valleys range between 10 and 60 Angstroms in height and depth. Correspondingly, the sharp fringes will not remain straight under increasing magnification but become so wiggly that one has difficulty in estimating their "center of gravity." These wiggly fringes are valuable for comparing different methods of polishing and molding the surfaces of optical elements. The method is so sensitive that the height of a molecule of mica has been determined with an accuracy that compares favorably with the result obtained from x-rays.

16.19 THE METHOD OF CHANNELED SPECTRA

16.19.1 General.

16.19.1.1 The conventional method for observing channeled spectra (also called the FECO bands, i.e., fringes of equal chromatic order) is illustrated in Figure 16.28 for the case in which the FECO bands are formed by transmission at the interferometer. Collimated white light passes through the interferometer. The image of the wedge is focused upon the entrance slit of a wavelength monochromator. One may view or photograph the FECO bands that appear at the exit pupil of the wavelength monochromator. If the surfaces S_1 and S_2 of the interferometer are plane, the interference bands seen at the eyepiece will be straight, of different wavelength, and of consecutive spectral order ν as indicated. In this method, the surfaces S_1 and S_2 are preferably parallel. Since it is an accidental matter to achieve parallelism by pressing surface S_1 against surface S_2, the practical compromise is to alter the relative inclination of surfaces S_1 and S_2 to the point at which the interference bands formed at the eyepiece of the wavelength monochromator are parallel to the image of the entrance slit.

16.19.1.2 The arrangement illustrated in Figure 16.26 allows white light to pass through the interferometer plates. Consequently, a relatively large amount of light flux is available to disturb the thermal equilibrium of the interferometer plates. The observed wavelengths of the interference bands can drift for hours before reliable readings can be taken. A more satisfactory arrangement that minimizes drifts due to thermal causes has been described by H. Osterberg and D. LaMarre.* Their arrangement, as applied to obtaining multiple beam fringes by reflection, is illustrated in Figure 16.29. Monochromatic light of measured, variable wavelength illuminates the interferometer. The interference fringes seen at the eyepiece of the microscope are of the same wavelength for a given setting of the wavelength drum and differ consecutively, as indicated, in order number. Indeed, the fringes resemble those of Figure 16.26 and could be measured as discussed in paragraph 16.18 with the aid of an eyepiece micrometer for determining the thickness of a film. To do so would defeat several advantages of this arrangement. Instead, advantage is taken of the fact that the fringes move as the wavelength drum is turned. In this way, consecutive fringes from each side of the step can be brought into coincidence with a fixed pointer or marker on the reticule and the corresponding wavelength recorded. With this arrangement the surfaces S_1 and S_2 should not be parallel but should be preferably (although not necessarily) inclined so that the multiple beam fringes are approximately perpendicular to the image of the step that marks the edge of the film whose thickness is to be measured. This step is imaged sharply upon the plane of the reticule. Consequently, each wavelength determination is made across a definite, localized, and selected area at the edge of the film. This area is that portion of the surface S_2 which is projected upon the pointer at the plane of the reticule. It follows that slight or even marked departures of the test surfaces from flatness have secondary effects upon the accuracy of this method of channeled spectra. One looks for a spot at which the fringe runs quite straight across the edge of the film and makes his measurements here.

16.19.1.3 The main advantage of the method of channeled spectra over the direct method of multiple beam fringes discussed in paragraph 16.18 is that the flatness of the surfaces S_1 and S_2 is much less critical for the purpose of making thickness measurements. A second advantage consists of the fact that channeled spectra enable one to measure either thin or thick films without ambiguity relative to whether the fringe shift exceeds or does not exceed a suspected number of fringe widths.

16.20 INTERPRETATION OF MEASUREMENTS WITH CHANNELED SPECTRA

16.20.1 Introduction.

16.20.1.1 Examination of the theory of multiple beam interferometry stated in paragraphs 16.15 through 16.17 shows that whether one is dealing with fringes obtained in either reflection or transmission from parallel plates or from wedges, the analytic condition for the appearance of the sharp fringes is of the form

$$\nu \lambda = 2d + \lambda f \; ; \quad n = 1 \; ; \qquad (131)$$

where ν is an integer, d is the separation of the interferometer surfaces, λ is the wavelength, and f is a function related to the phase changes that take place on reflection at the coated surfaces of the reflecting surfaces. The function f can vary with wavelength and will be different for the transmitted and reflected fringes.

---

*H. Osterberg and D. LaMarre, J. Opt. Soc. Amer., 46, 777-778 (1956).

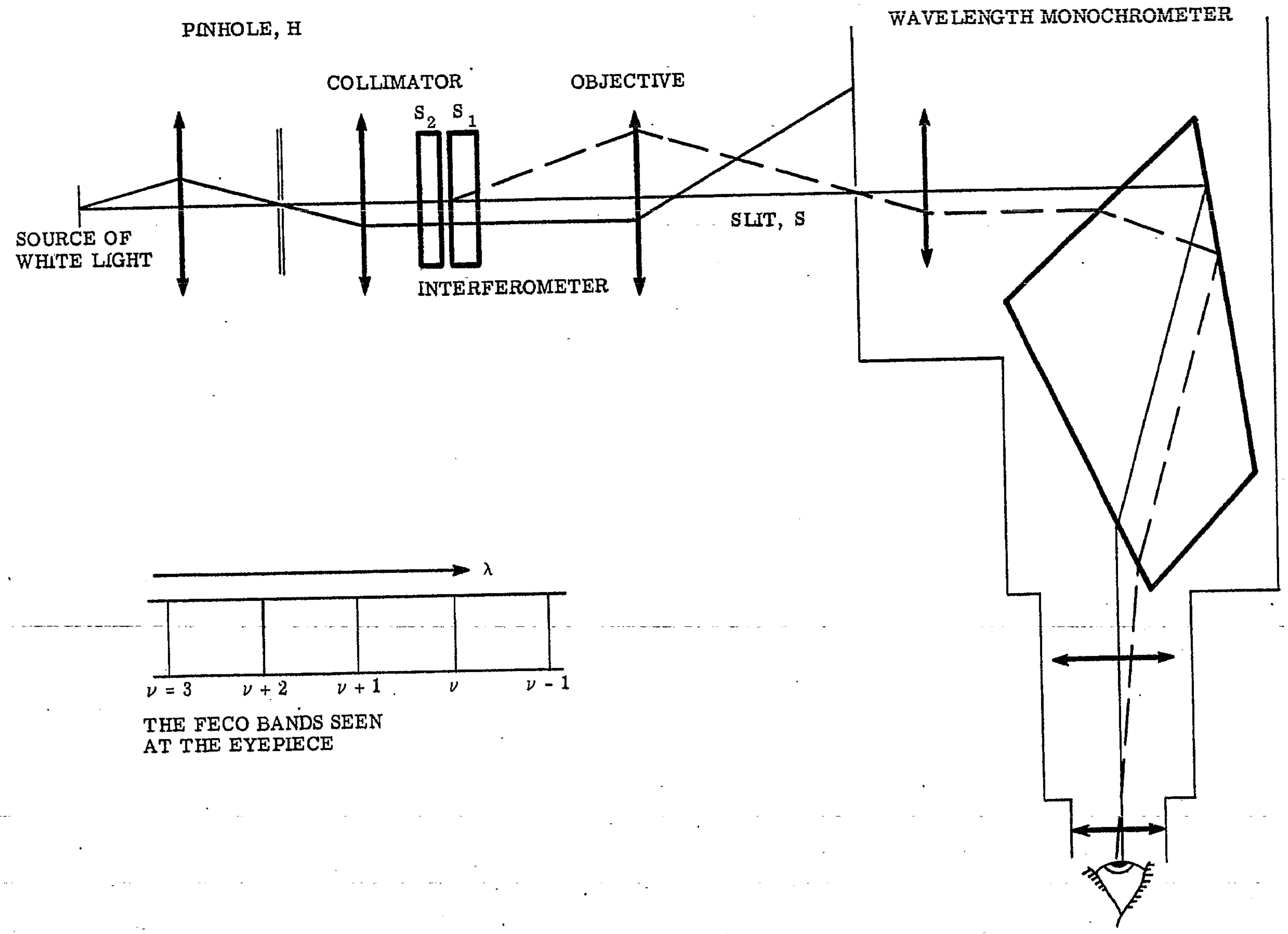


FIGURE 16. 28- Conventional method for obtaining channeled spectra from a source of white light.

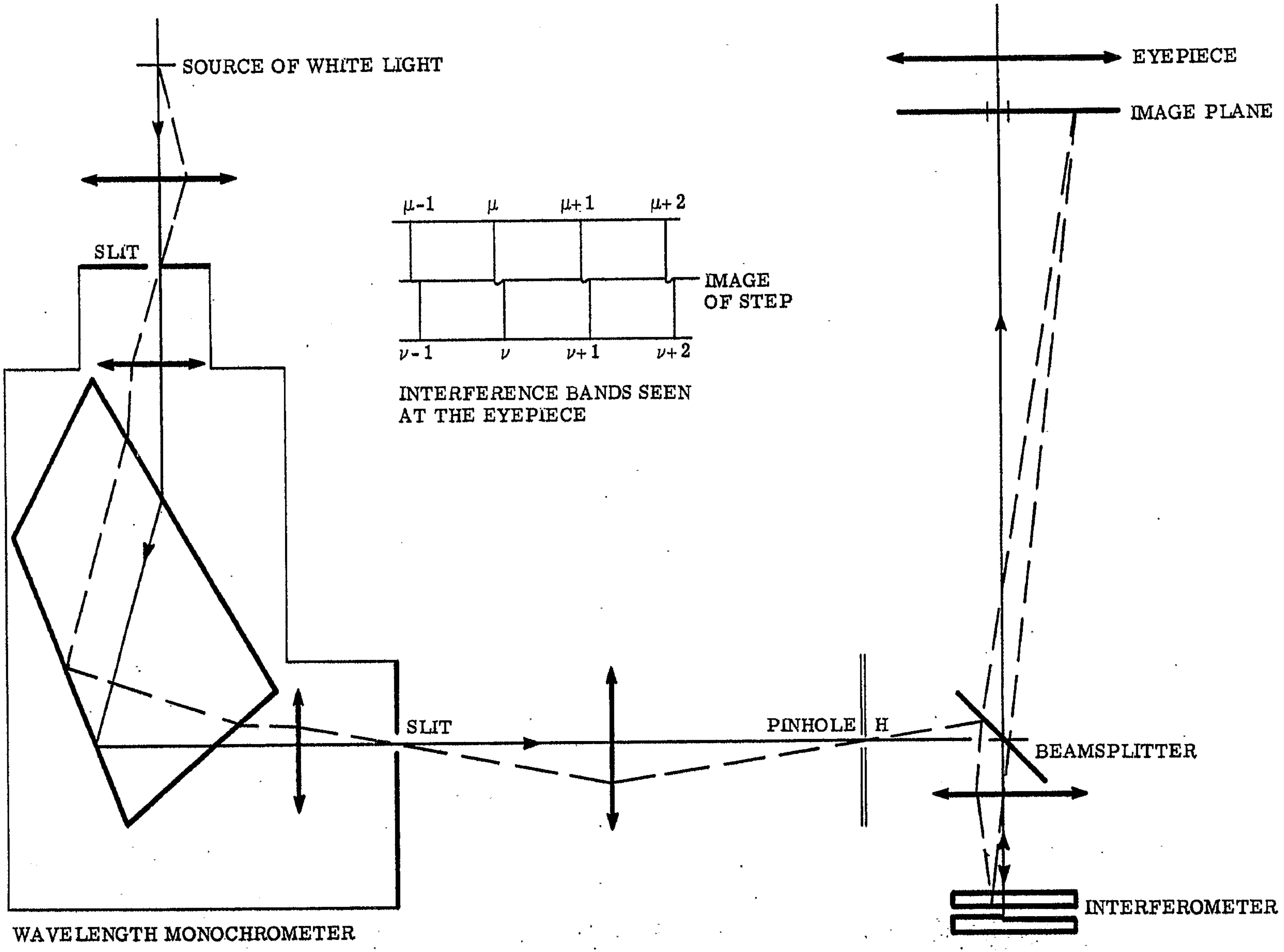


FIGURE 16. 29–A second method for producing channeled spectra.

16.20.1.2 With regard to the transmitted fringes, it has been customary to take $f = 0$ and to state that each bright fringe occurs at those wavelengths $\lambda = \lambda_\nu$ for which $d = \nu(\lambda\nu/2)$. Interpretations based upon this simplified view are, however, inadequate.

16.20.1.3 Let f be expanded as a function of wavelength about the wavelength λ_o such that

$$f = f_o + b(\lambda - \lambda_o) + c(\lambda - \lambda_o)^2 + \dots. \qquad (132)$$

Experience has shown that with silver coatings or with high reflecting multilayers, there will exist an extended range for which $\nu\lambda$ is, with good approximation, a linear function of λ about an appropriately chosen λ_o in the visible region. For this range of wavelengths the first two terms of Taylor's expansion of $\nu\lambda$ about the point $\lambda = \lambda_o$ from Equations (131) and (132) yield the approximation

$$\nu\lambda = 2d - b\lambda_o^2 + s_o\lambda \qquad (133)$$

in which

$$s_o = f_o + b\lambda_o. \qquad (134)$$

16.20.1.4 With respect to Figure 16. 28, consider two nearby spectral orders ν and $\nu + p$ where $p = 0, \pm 1, \pm 2, \pm 3$, etc. Let λ_ν and $\lambda_{\nu+p}$ be the central wavelengths of these spectral bands. We have seen that channeled spectra are obtained from a single localized area for which the separation d of the interferometer mirrors is constant. Since $b\lambda_o^2$ is constant, it follows from Equation (133) that

$$\lambda_\nu(\nu - s_o) = 2d - b\lambda_o^2 = \text{constant} = \lambda_{\nu+p}(\nu + p - s_o). \qquad (135)$$

Hence,

$$\nu - s_o = p\,\frac{\lambda_{\nu+p}}{\lambda_\nu - \lambda_{\nu+p}}. \qquad (136)$$

As will be seen from Figure 16. 28, determining p is simply a matter of counting bands from the band whose order number is labelled ν. Since p, λ_ν, and $\lambda_{\nu+p}$ are known, one can compute $\nu - s_o$ from Equation (136). It is good practice to compute $\nu - s_o$ for at least three values of $|p|$ when enough bands are available. If the values $\nu - s_o$ thus obtained are not alike within a range corresponding to one's experimental error in reading the wavelengths, the separation d of the interferometer mirrors is changing or $|p|$ has been chosen so large that $\lambda_{\nu+p}$ falls outside of the range for which $\nu\lambda$ is adequately linear in λ.

16.20.1.5 The values $\nu - s_o$ will fall in the range 10 to 70 when the interferometer mirrors are laid one upon the other except when great precaution is taken to avoid dust particles. The corresponding separation of the interferometer mirrors falls in the range 5 to 35 wavelengths. This explains why the separation can vary with temperature, etc. When $\nu - s_o$ has been determined, the separation d is given by

$$2d = (\nu - s_o)\lambda_\nu + b\lambda_o^2. \qquad (137)$$

Unfortunately, one needs to know $b\lambda_o^2$ in order to compute d accurately. One may, of course, accept $2d = (\nu - s_o)\lambda_\nu$ as his approximation and expect that this will be a better approximation than obtained by asserting that $2d = \nu\lambda_\nu$.

16.20.1.6 On the other hand, a knowledge of $b\lambda_o^2$ is not required in order to determine accurately the thickness t of a film. With respect to the interference bands $\nu + p$ and $\mu + p$ seen on each side of the step formed at the edge of the film (see Figure 16. 29), one determines $\nu - s_o$ and $\mu - s_o$ from the wavelengths λ_ν, $\lambda_{\nu+p}$ and λ_μ, $\lambda_{\mu+p}$ for which the interference bands are brought into coincidence with the marker on the reticule by turning the wavelength drum. The "non-integral spectral orders" $\nu - s_o$ and $\mu - s_o$ become known on each side of the step at the film. Then, from Equation (137)

$$2t = 2(d_1 - d_2) = (\nu - s_o)\lambda_\nu - (\mu - s_o)\lambda_\mu. \qquad (138)$$

If the film is thin enough, one finds automatically that $\nu - s_o = \mu - s_o$ or that $\nu = \mu$. In such cases Equation (138) reduces to

$$t = \frac{1}{2}(\nu - s_o)(\lambda_\nu - \lambda). \qquad (139)$$

16.20.1.7 Let us consider the sensitivity and accuracy of the method of channeled spectra in, for example, the measurement of the thickness t of the thin films to which Equation (139) applies. If the error in reading the wavelengths λ_ν and λ_μ is $\delta\lambda$ and if δt is the corresponding error in t, then for estimating δt, we observe from Equation (139) that

$$\delta t \leqq \left(\frac{\nu - s_o}{2}\right) 2 \;|\delta\lambda| \;\leqq\; (\nu - s_o)\;|\delta\lambda| \;. \qquad (140)$$

It becomes clear that the error $|\delta t|$ is reduced by making measurements at low values of $\nu - s_o$, i.e., at low separations d of the interferometer mirrors. Reducing $\nu - s_o$ to values in the neighborhood of 1 or 2 causes the spectral bands to broaden and to become excessively wiggly when polished surfaces are employed. The added difficulty of setting upon the center of gravity of the interference bands now appears. With the use of diffraction gratings, such as monochromators, and of photographic methods involving microdensitometry, errors $|\delta t|$ of 0.1 Angstrom or less may become possible. To carry the method to such extremes is however costly, cumbersome, and tedious. A typical example of the actual error obtained by making routine visual settings with a prism monochromator has been cited by Osterberg and LaMarre. They found that the visual settings with a Hilger Barfit monochromator are reproducible to about one Angstrom. With $\nu - s_o = 35$, the corresponding maximum error δt in the thickness t of the film is 35 Angstroms. The actual computed values of t from a series of spectral orders $\nu + p$ and $\mu + p$ agree to about 10 Angstroms.

16.20.1.8 One should not form the impression that the method of channeled spectra is restricted to analysis of fringes produced by multiple beam interferometry. We have seen, for example, that order numbers ν are associated with Fizeau fringes as in Equation (61). By projecting Fizeau fringes formed in white light upon the slit of a wavelength monochromator as shown in Figure 16.29 or by adapting the modification illustrated in Figure 16.29, a series of bands will be seen at the eyepiece. Comparison of Equations (61) and (131) shows that one deals with the simpler case f = 0 in applying the method of channeled spectra to Fizeau fringes.

16.21 HUYGENS' PRINCIPLE

16.21.1 Introduction. Although Huygens' principle is less general than, for example, Kirchhoff's law, its applications are far simpler to follow and yield predictions that are in resonable close accord with experiment with respect to the phenomena that we shall consider.

16.21.1.1 Huygens' principle supposes that as a wave travels through a homogeneous, isotropic space, each point in the space is excited as the wave passes through it and serves as origin for a spherical wave that expands with the velocity of light in the medium. Requirements such as conservation of energy require that the amplitude of the spherical wave decrease as 1/r where the distance r is measured from the point of expansion. Furthermore, the principle supposes that the propagation of the wave itself through space is a consequence of the interference effects that take place between the infinite set of expanding spherical wavelets. Close examination of this interference process shows, for example, that the reconstructed wave thus obtained from an assumed plane wave travelling to the right is, in turn, a plane wave that travels to the right. The wave that tends to travel to the left is destroyed, in effect, by destructive interference. The development of a wavefront as the envelope of the spherical wavelets that expand from the original wavefront at $z = z_o$ is illustrated in Figure 16.30

16.21.1.2 The construction of Figure 16.31 enables one to deduce Snell's law of refraction from Huygen's principle. If t_o is the time required for light to travel from C to B in the 0^{th} medium,

$$CB = v_o\, t_o = \frac{c t_o}{n_o} \;.$$

The spherical wave starting from A travels the distance AD in time t_o such that

$$AD = v_1\, t_o = \frac{c t_o}{n_1} \;.$$

But

$$\sin i = \frac{CB}{AB} \quad ; \quad \sin i' = \frac{AD}{AB} \;.$$

Hence,

$$\frac{\sin i}{\sin i'} = \frac{CB}{AD} = \frac{n_1}{n_o}$$

This demonstration shows that the most basic law of geometrical optics can be explained by diffraction.

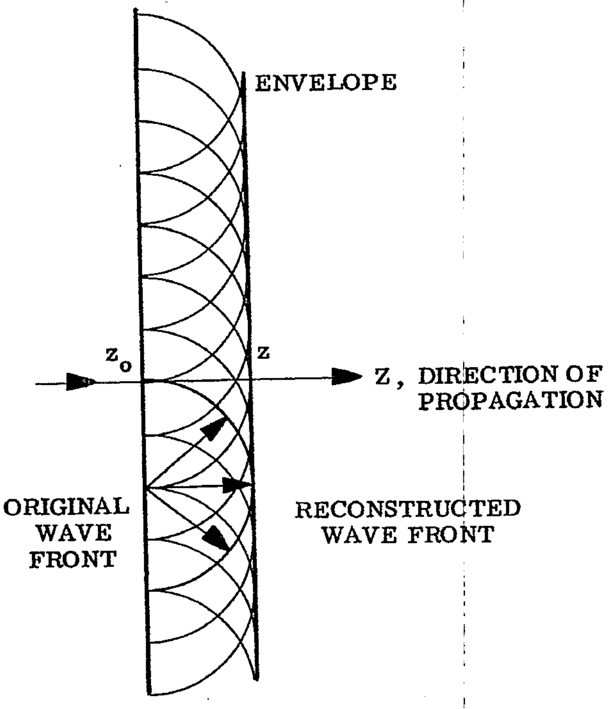


FIGURE 16. 30- Propagation of a plane wave in accordance with Huygens' Principle.

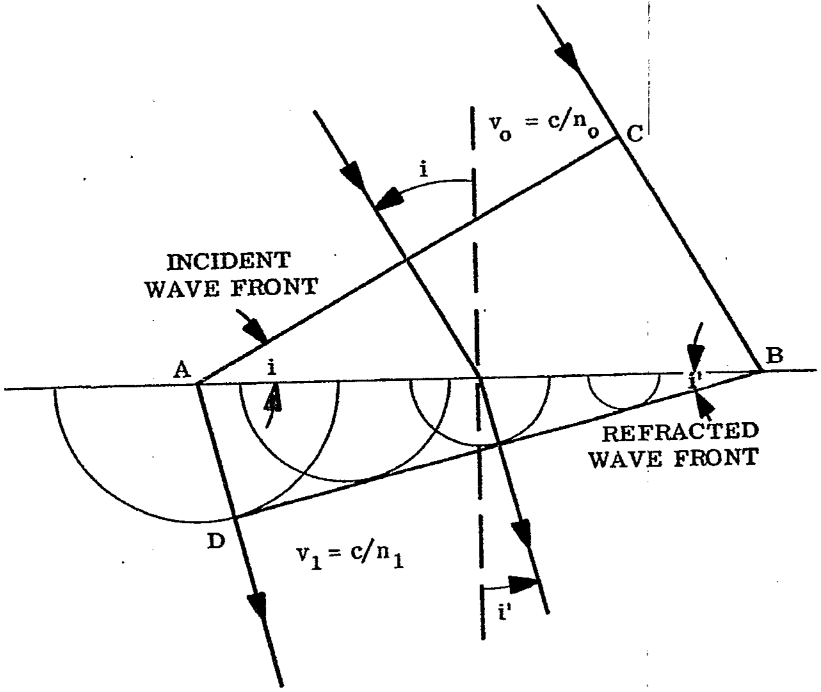


FIGURE 16. 31- Construction for obtaining Snell's Law of Refraction from Huygens' Principle.

16.21.1.3 We shall need an analytical statement of Huygens' principle. The amplitude and phase variation of the electric vector of a spherical wave that expands from any point O in a space whose refractive index is n is given by

$$E = \frac{e^{i(knr - \omega t)}}{r} \qquad (141)$$

where $k \equiv 2\pi/\lambda$; $\omega \equiv 2\pi/T$; and r is distance measured from point O. The physical meaning of Equation (141) is, of course, in doubt at the point $r = 0$ - but not elsewhere.

16.22 FRAUNHOFER DIFFRACTION

16.22.1 Discussion of theory.

16.22.1.1 Fortunately, the theory and interpretation of diffraction phenomena become much simpler when these phenomena are considered at relatively large distances from the diffracting aperture or obstacle. When a lens is placed between the aperture and the plane at infinity, the diffraction phenomena at infinity are brought into the focal plane of the lens. This consideration leads one to suspect that diffraction phenomena that occur at the focal plane of lenses are likely to be Fraunhofer diffraction phenomena. Since diffraction effects associated with focal planes belong to the classification known as Fraunhofer diffraction phenomena, these diffraction phenomena are of primary fundamental interest to the designer of optical (or radar) instruments.

16.22.1.2 Simplified arguments based upon Huygens' construction can be used to locate maxima and minima in the energy densities associated with Fraunhofer diffraction effects, but such arguments do not predict the distribution of energy density. The following diffraction integrals become so simple and direct that we shall omit the elementary and less instructive theory. The diffraction integral governing Fraunhofer diffraction is easily integrated or applied to a large number of practical cases.

16.22.1.3 We suppose that the aperture or obstacle from which diffraction occurs is located at the $\zeta\eta$ plane of Figure 16.32 and that the observation plane xy is located at distance D from the $\zeta\eta$ plane. Huygens wavelets leave each element of area $d\zeta\, d\eta$ of the $\zeta\eta$ plane and arrive at point P of the plane of observation after traversing the distance r where

$$r = \left[(x - \zeta)^2 + (y - \eta)^2 + D^2\right]^{1/2} \qquad (142)$$

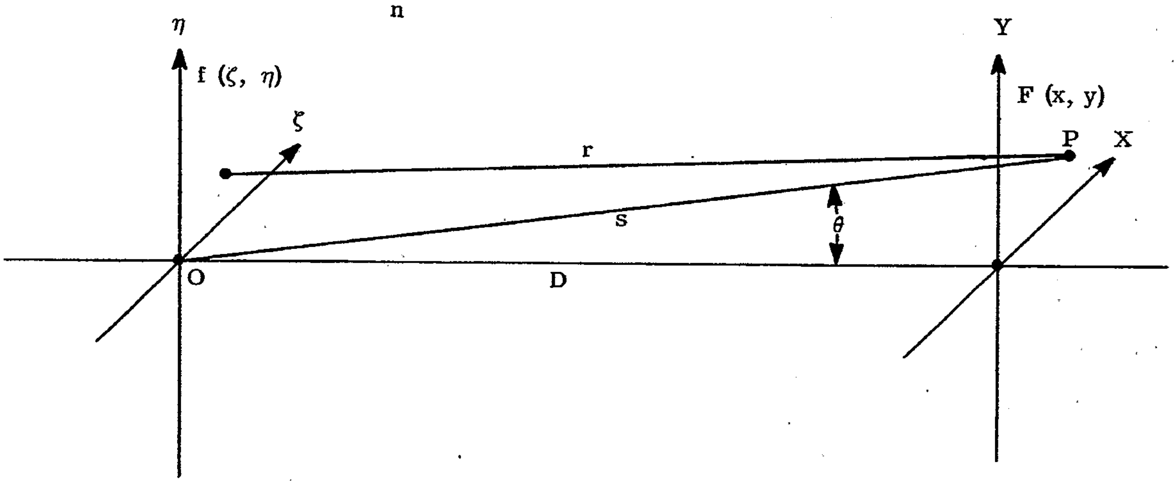


FIGURE 16.32 -Convention with respect to the integral statement of Huygens' Principle.

These Huygens wavelets expand from point (ζ, η) as described by Equations (141) and (142). Our problem is to sum the Huygens wavelets that leave all points (ζ, η) of the plane of the aperture and arrive at point P.

16.22.1.4 To formulate the problem a bit more generally without adding unduly to the complexity of presenting the problem, we can suppose that $f(\zeta, \eta)\, d\zeta\, d\eta$ is a complex number that specifies the amplitude and phase of the coherent Huygens wavelets that leave the area $d\zeta\, d\eta$. (We shall deal mainly with the simple cases in which $f(\zeta, \eta) = 1$.) According to Equation (141), the Huygens wavelets that leave the area $d\zeta\, d\eta$ with the amplitude and phase expressed by $f(\zeta, \eta)\, d\zeta\, d\eta$ arrive at point P with the amplitude and phase given by

$$f(\zeta, \eta)\, d\zeta\, d\eta\, \frac{e^{i(knr - \omega t)}}{r} .$$

Let $F(x, y)$ be the complex number that denotes the sum of all of the interfering Huygens wavelets that arrive at the point of observation P of Figure 16. 32. From the theory of integral calculus this sum is given at once by the integral

$$F(x,y) = e^{-i\omega t} \int\int f(\zeta, \eta)\, \frac{e^{iknr}}{r}\, d\zeta\, d\eta \tag{143}$$

in which the integration extends over the illuminated area of the $\zeta\eta$ plane and in which r is given by Equation (142).

16.22.1.5 Before passing on to the Fraunhofer form of the integral given in Equation (143), we remark that the term Fresnel diffraction (as distinguished from Fraunhoffer diffraction) is applied to the cases in which the distance D from the plane of the aperture to the plane of observation is relatively small. Equation (143) is the most general statement of Huygens' principle. It includes Fresnel and Fraunhofer diffraction as special cases.

16.22.1.6 The Fraunhofer specialization diffraction integral is obtained in the following way from the integral of Equation (143) and the supposition that D is large. By expanding the squares in Equation (142) and defining

$$S \equiv \left[D^2 + x^2 + y^2\right]^{1/2} , \tag{144}$$

one finds that

$$r = S\left[1 + \frac{\zeta^2 + \eta^2}{S^2} - \frac{2(x\zeta + y\eta)}{S^2}\right]^{1/2} \tag{145}$$

in which S has the geometrical meaning illustrated in Figure 16. 32. We suppose that D becomes great but that the aperture opening at the $\zeta\eta$ plane remains finite. Equivalently, but somewhat more generally, we may say that $f(\zeta, \eta) = 0$ when $(\zeta^2 + \eta^2)^{1/2}$ exceeds some finite value and that D can approach infinity. Under these circumstances, the quantity $(\zeta^2 + \eta^2)/S^2$ in Equation (145) is surely negligible. Since x and y can become infinite at $D = \infty$, the quantity $2(x\zeta + y\eta)/S^2$ is not entirely negligible. Because ζ and η will be small in comparison to S, $1 >> 2(x\zeta + y\eta)/S^2$. Hence, with excellent approximation,

$$\left[1 - 2\left(\frac{x\zeta + y\eta}{S^2}\right)\right]^{1/2} = 1 - \frac{x\zeta + y\eta}{S^2} \tag{146}$$

Therefore,

$$r = S - \frac{x\zeta + y\eta}{S} \tag{147}$$

16.22.1.7 Upon introducing r from Equation (147) into Equation (143), it will suffice to set $r = S$ in the denominator since $(x\zeta + y\eta)/S$ will be very small. However, the quantity $(x\zeta + y\eta)/S$ is multiplied by the large factor $k = 2\pi/\lambda$ in the exponent. We now introduce $r = S$ in the denominator of Equation (143) and r from Equation (147) into the exponent and thus obtain

$$F(x, y) = e^{-i\omega t}\, \frac{e^{iknS}}{S} \int\int f(\zeta, \eta)\, e^{-ikn\frac{x\zeta + y\eta}{S}}\, d\zeta\, d\eta \tag{148}$$

in which

$$\begin{aligned} S &= (D^2 + x^2 + y^2)^{1/2} \\ k &= 2\pi/\lambda \\ \omega &= 2\pi/T . \end{aligned} \tag{149}$$

$f(\zeta, \eta)$ specifies the amplitude and phase of the disturbance as it leaves the plane of the aperture. The integration extends over the plane of the aperture. In case the aperture consists, for example, of an opaque screen with a hole in it, the integration with respect to $d\zeta\, d\eta$ extends over the area of the hole. $F(x, y)$ is a complex number that specifies the amplitude and phase of the so-called Fraunhofer region.

16.22.1.8 The energy density, $W(x, y)$, is proportional to $|F(x, y)|^2$. Since $|e^{-i\omega t}|^2 = 1$ and $|e^{iknS}|^2 = 1$, it follows from Equation (148) that

$$W(x, y) = \frac{1}{S^2}\, |F_o(x, y)|^2 \qquad (150)$$

where

$$F_o(x, y) = \underset{\text{over plane of aperture}}{\int\int} f(\zeta, \eta)\, e^{-ikn\frac{x\zeta + y\eta}{S}}\, d\zeta\, d\eta\, . \qquad (151)$$

It suffices therefore to compute the slightly simpler integral, $F_o(x, y)$, of Equation (151) when one wishes to determine the time-averaged distribution $W(x, y)$ of energy density produced at point (x, y) by the radiation in a coherent wave that illuminates the $\zeta\eta$ plane of the aperture.

16.23 FRAUNHOFER DIFFRACTION FROM A RECTANGULAR APERTURE

16.23.1 Discussion of principles.

16.23.1.1 We suppose for simplicity that the rectangular aperture is illuminated as in Figure 16.32 by a plane wave at normal incidence. It suffices to set

$$f(\zeta, \eta) = \text{constant} = 1\, . \qquad (152)$$

Then, from Equation (151),

$$\begin{aligned} F_o(x, y) &= \int_{-a}^{a}\int_{-b}^{b} e^{\frac{-iknx\zeta}{S}}\, e^{\frac{-ikny\eta}{S}}\, d\zeta\, d\eta \\ &= \int_{-a}^{a} e^{\frac{-iknx\zeta}{S}}\, d\zeta \int_{-b}^{b} e^{\frac{-ikny\eta}{S}}\, d\eta \\ &= \frac{e^{\frac{iknxa}{S}} - e^{\frac{-iknyb}{S}}}{(iknx)/S} \cdot \frac{e^{\frac{iknxa}{S}} - e^{\frac{-iknyb}{S}}}{(ikny)/S}\, . \end{aligned}$$

Since $k = 2\pi/\lambda$ and $\sin z = (e^{iz} - e^{-iz})/2i$,

$$F_o(x, y) = 4ab \left[\frac{\sin(2\pi anx/S\lambda)}{2\pi anx/S\lambda}\right]\left[\frac{\sin(2\pi bny/S\lambda)}{2\pi bny/S\lambda}\right]\, . \qquad (153)$$

From Equations (152) and (150), the corresponding time-averaged distribution of energy density in the observation plane is given by

$$W(x, y) = \frac{16a^2b^2}{S^2}\left[\frac{\sin(2\pi anx/S\lambda)}{2\pi anx/S\lambda}\right]^2\left[\frac{\sin(2\pi bny/S\lambda)}{2\pi bny/S\lambda}\right]^2 . \qquad (154)$$

Along, for example, the line $y = 0$,

$$W(x, 0) \equiv W(x) = \frac{A^2}{S^2}\left[\frac{\sin(2\pi anx/S\lambda)}{2\pi anx/S\lambda}\right]^2 \qquad (155)$$

because $(\sin u)/u = 1$ when $u = 0$. $A \equiv 4ab$ is the area of the rectangular aperture. We can take $S \cong D$ for most purposes. $W(x)$ assumes its greatest value $W = A^2/S^2$ at $x = 0$. $W(x)$ decreases as $1/x^2$. The energy density is zero whenever $(2\pi anx)/S\lambda = \nu\pi$ where ν is an integer. Hence, the zeros of $W(x)$ occur at the points x for which

$$\frac{x_\nu}{S} = \sin\theta_\nu = \frac{\nu\lambda}{2an} \qquad (156)$$

where $\nu = \pm 1, \pm 2, \pm 3$, etc.; 2a is the width of the rectangular aperture along the x-direction; n is the refractive index of the space; and θ_ν is the angle θ (Figure 16.33) that corresponds to x_ν along the line $y = 0$.

16.23.1.2 Similiar conclusions hold along the line $x = 0$. One has only to substitute y for x and b for a in Equations (155) and (156). The width of the aperture along the y-direction is 2b.

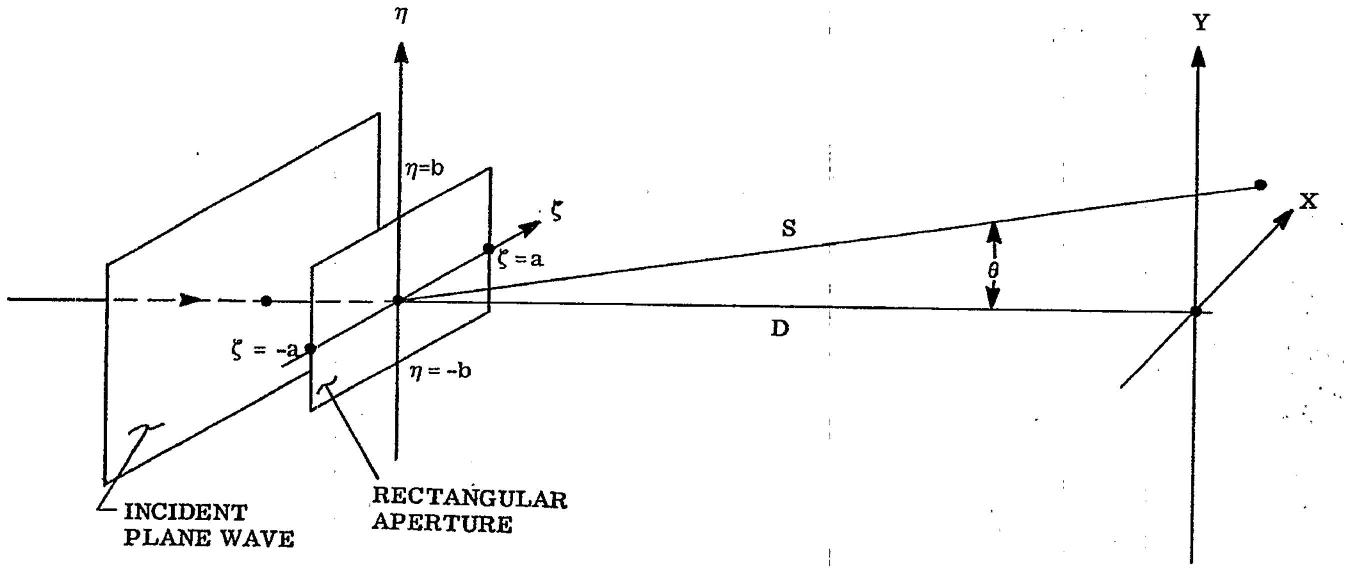


FIGURE 16.33-Notation with respect to diffraction from a rectangular aperture illuminated at normal incidence.

16.24 FRAUNHOFER DIFFRACTION FROM CIRCULAR APERTURES

16.24.1 Discussion of principles.

16.24.1.1 We suppose that a plane wave is incident normally upon the circular aperture so that $f(\zeta, \eta) = 1$. It is convenient to replace ζ, η and x, y by polar coordinates because the aperture is circular. Let

$$\zeta = u \cos \phi \quad ; \quad \eta = u \sin \phi \quad ; \tag{157}$$

$$x = r \cos \alpha \quad ; \quad y = r \sin \alpha \quad ; \tag{158}$$

in which the geometrical meanings of u, ϕ, r, and α are illustrated in Figure 16.34. Upon introducing Equations (157) and (158) into Equation (151) and setting $f(\zeta, \eta) = 1$, one obtains

$$F_o(x, y) \equiv F_o(r) = \int_o^a \int_o^{2\pi} e^{\frac{-iknru}{S}\cos(\phi-\alpha)} \quad u\,du\,d\phi \tag{159}$$

in which

$$S = (D^2 + x^2 + y^2)^{1/2} = (D^2 + r^2)^{1/2}. \tag{160}$$

16.24.1.2 One can prove that $F_o(r)$ must be independent of α because the integrand of Equation (159) is periodic in the angle ϕ. However, it is clear from Figure 16. 34 that $F_o(r)$ should be independent of the angle α because the system has complete axial symmetry. Hence, we can set

$$\alpha = 0 \tag{161}$$

in Equation (159).

16.24.1.3 Now one finds from almost all text books treating the elementary theory of Bessel functions that

$$\int_o^{2\pi} e^{\pm iz\cos\phi} \quad d\phi = 2\pi J_o(z) \tag{162}$$

where $J_o(z)$ is a Bessel function of zero order and first kind. From Equations (159) and (162)

$$F_o(r) = 2\pi \int_o^a J_o\left(\frac{knru}{S}\right) u\,du. \tag{163}$$

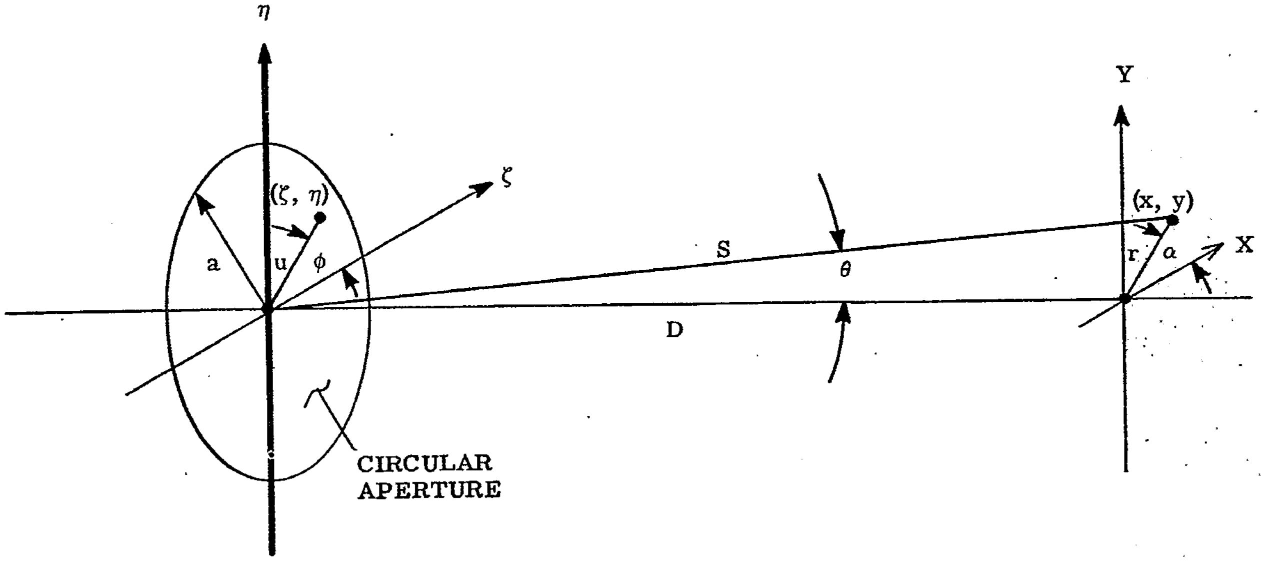


FIGURE 16. 34-Notation with respect to Fraunhofer Diffraction from circular apertures.

Introduce the change of variable

$$v \equiv \frac{knru}{S} \quad , \text{ or } \quad u \equiv \frac{S}{knr} v \ . \tag{164}$$

Then,

$$F_o(r) \equiv 2\pi \left(\frac{S}{knr} \right)^2 \int_0^{\frac{knra}{S}} v \, J_o(v) \, dv \ . \tag{165}$$

It is another elementary proposition in Bessel Functions that

$$\int_0^z v \, J_o(v) \, dv = z \, J_1(z) \tag{166}$$

where $J_1(z)$ is a Bessel function of first order and first kind. Since $J_1(z) = 0$ at $z = 0$, one finds directly from Equations (165) and (166) that

$$F_o(r) = 2\pi \left(\frac{S}{knr} \right)^2 \frac{knra}{S} \ J_1 \left(\frac{knra}{S} \right) \ .$$

Whence

$$F_o(r) = 2\pi a^2 \ \frac{J_1(2\pi na \ r/\lambda S)}{(2\pi nar)/\lambda S} \ . \tag{167}$$

16.24.1.4 We note from Figure 16.34 that

$$\sin\theta = r/S \ . \tag{168}$$

Alternatively, one may therefore write

$$F_o(\theta) \equiv 2\pi a^2 \cdot \frac{J_1(2\pi na \sin\theta/\lambda)}{2\pi na \sin\theta/\lambda} \tag{169}$$

The energy density in the Fraunhofer diffraction image or pattern from a circular aperture of radius $\underline{a}$ is now given by Equation (150) in which one introduces F_o from Equation (167) or from Equation (169).

16.24.1.5 The function $[J_1(z)]/z$ is 1/2 at $z = 0$ and assumes its first zero at $z = 3.8317$. Therefore, the energy density in the Fraunhofer diffraction pattern has its first zero minimum at $(2\pi na \sin\theta_1)/\lambda = 3.8317$ or at

$$\sin\theta_1 = \frac{r_1}{S} = \frac{0.61\lambda}{na} = \frac{1.22\lambda}{2an} \tag{170}$$

in which 2a is the diameter of the aperture. It is instructive to compare Equations (170) and (156) at the first zero where $\nu = 1$. We see that the central maximum in the diffraction pattern is 22 per cent larger in linear dimension for the circular aperture than for the rectangular aperture whose width is equal to the diameter of the circular aperture. The Bessel function $J_1(z)$ oscillates with increasing z in such a way that successive maxima and minima of $J_1(z)$ decrease numerically. Hence, the energy density

$$W(r) = \frac{4\pi^2 a^4}{S^2}\left[\frac{J_1(2\pi nar/S\lambda)}{2\pi nar/S\lambda}\right]^2 \tag{171}$$

in the diffraction pattern produced by a circular aperture decreases considerably faster with increasing distance r from the diffraction head than does the energy density W (x) produced by a rectangular aperture. (Compare Equations (155) and (171).) One must expect that circular apertures are preferable to rectangular apertures for lenses because the diffraction images produced by circular apertures are, on the whole, more concentrated.

16.25 DIFFRACTION FROM SPHERICAL WAVEFRONTS

16.25.1 General. Whereas the methods of paragraphs 16.23 and 16.24 can be utilized as a basis for discussing the diffraction images produced by lenses, the adaptation of these methods is a bit too artificial and leads, awkwardly, to the predictions that resolving power is related to the tangent of certain axial angles rather than to the sine of these angles.

16.25.1.1 It is the purpose of a well corrected lens to convert a spherical wave that diverges from an object point into a spherical wave that converges upon the conjugate image point as in Figure 16. 34. We suppose for

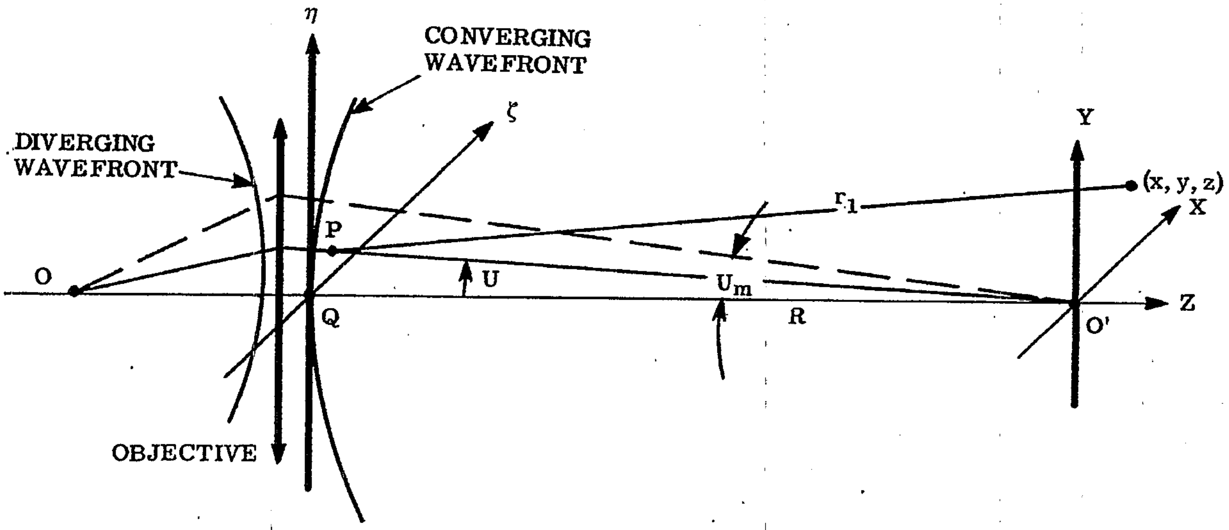


FIGURE 16. 35- Convention with respect to the formation of a diffraction image, O', of O by a lens system.

simplicity of presentation that the object point O is located upon the axis. Let V be the optical path from O to O'. We draw a reference sphere of radius R about the point O' such that this sphere touches the tangent plane $\zeta\eta$ at point Q on the axis. The optical path from O to Q is now V - nR where n is the refractive index of the image space. Similarly, the optical path from point O to any point P on the $\zeta\eta$ plane from point O to any point P on the $\zeta\eta$ plane is

$$V - n\,O'\,P = V - n\,(R^2 + \zeta^2 + \eta^2)^{1/2} \tag{172}$$

in the absence of spherical aberration. The Huygens wavelets now leave the $\zeta\eta$ plane with an amplitude-phase distribution given by

$$f(\zeta, \eta) = \frac{e^{ikV}\, e^{-ikn\,(R^2 + \zeta^2 + \eta^2)^{1/2}}}{(R^2 + \zeta^2 + \eta^2)^{1/2}} \tag{173}$$

16.25.1.2 We choose the origin of the coordinates X, Y, Z at the point O' with O' conjugate to O. Thus, the plane z = 0 is the sharply focused image plane. The problem is to find the amplitude-phase distribution F (x, y, z) produced by all the Huygens wavelets that leave the $\zeta\eta$ plane. From Equations (143) and (173),

$$F(x,y,z) = e^{-i\omega t}\, e^{ikV} \iint \frac{e^{-ikn\,(R^2 + \zeta^2 + \eta^2)^{1/2}}}{(R^2 + \zeta^2 + \eta^2)^{1/2}} \cdot \frac{e^{iknr_1}}{r_1}\, d\zeta\, d\eta \tag{174}$$

where the distance r_1 of Figure 16. 35 is

$$r_1 = \quad (x - \zeta)^2 + (y - \eta)^2 + (R + z)^2 \quad {}^{1/2} \tag{175}$$

However, one finds after slight rearrangement that

$$r_1 = (R^2 + \zeta^2 + \eta^2)^{1/2} \left[\frac{1 + x^2 + y^2 + z^2 - 2\,(x\zeta + y\eta - Rz)}{R^2 + \zeta^2 + \eta^2} \right]^{1/2} \tag{176}$$

16.25.1.3 In order to obtain the approximation to r_1 that leads to the conventional diffraction integral for lenses, we have to suppose that the field of view is so small that one can afford to neglect the term $(x^2 + y^2 + z^2) / (R^2 + \zeta^2 \eta^2)$ in Equation (176). This means that the following theory holds best for small fields of view. We have to suppose also that the dimensions of the aperture at the $\zeta\eta$ plane and the out-of-focus distance z are small enough for us to be willing to accept the approximation

$$\left[1 - 2\, \frac{(x\zeta + y\eta - Rz)}{R^2 + \zeta^2 + \eta^2} \right]^{1/2} = 1 - \frac{x\zeta + y\eta - Rz}{R^2 + \zeta^2 + \eta^2} \tag{177}$$

Under these approximations,

$$r_1 = (R^2 + \zeta^2 + \eta^2)^{1/2} - \frac{x\zeta + y\eta - Rz}{(R^2 + \zeta^2 + \eta^2)} \quad . \tag{178}$$

16.25.1.4 Upon substituting r_1 from Equation (178) into the integral (174) it suffices to set $r_1 = (R^2 + \zeta^2 + \eta^2)^{1/2}$ in the denominator. Our approximation for F (x, y, z) becomes

$$F(x, y, z) = e^{-i\omega t}\, e^{iknV} \iint \frac{e^{-ikn \frac{x\zeta + y\eta - Rz}{\sqrt{R^2 + \zeta^2 + \eta^2}}}}{R^2 + \zeta^2 + \eta^2}\, d\zeta\, d\eta \tag{179}$$

in which the integration extends over the aperture of the objective, Figure 16. 35.

16.25.1.5 Since $F|(x, y, z)|^2$ is independent of ωt and V, it is convenient to drop the external exponentials in Equation (179) and to write again

$$F_o(x, y, z) \equiv \iint \frac{e^{-ikn \frac{x\zeta + y\eta - Rz}{\sqrt{R^2 + \zeta^2 + \eta^2}}}}{R^2 + \xi^2 + \eta^2}\, d\zeta\, d\eta \quad . \tag{180}$$

The time-averaged energy density in the diffraction image of an object point located upon the axis is

$$W(x, y, z) = |F(x, y, z)|^2 = |F_o(x, y, z)|^2 \quad . \tag{181}$$

The plane z = 0 is the sharply focused image plane.

16.26 PRIMARY DIFFRACTION INTEGRALS WITH OBJECTIVES HAVING CIRCULAR APERTURES

16.26.1 Introduction. We shall call the integral $F_o(x, y, z)$ of Equation (180) the primary diffraction integral and shall refer to the corresponding distribution of energy density $W(x, y, z)$ as the primary diffraction image. These two quantities are of fundamental importance to the diffraction theory of optical instruments. In this section, the primary diffraction integral will be specialized to the great class of objectives that have circular apertures. Thus far, the objective has been assumed free of spherical aberration.

16.26.1.1 Corresponding elements of Figures 16. 35 and 16. 36 are labeled alike. One notes from Figure 16. 36 that

$$\zeta = R \tan U \cos \phi \quad ; \quad \eta = R \tan U \sin \phi \quad ; \qquad (182)$$

and that

$$\cos U = R \,/\, \sqrt{R^2 + \zeta^2 + \eta^2} \ . \qquad (183)$$

Hence,

$$\frac{\zeta}{\sqrt{\zeta^2 + \eta^2 + R^2}} = \sin U \cos \phi \quad ; \quad \frac{\eta}{\sqrt{\zeta^2 + \eta^2 + R^2}} = \sin U \sin \phi \ . \qquad (184)$$

16.26.1.2 It is convenient to change the variables of integration from ζ and η to U and ϕ. Since

$$dA = \begin{vmatrix} \frac{\partial \zeta}{\partial U} & \frac{\partial \eta}{\partial U} \\ \frac{\partial \zeta}{\partial \phi} & \frac{\partial \eta}{\partial \phi} \end{vmatrix} dU\, d\phi = d\zeta\, d\eta \ ,$$

wherein ζ and η are given by Equation (182)

$$d\zeta\, d\eta = R^2 \frac{\sin U}{\cos^3 U}\, dU\, d\phi \ . \qquad (185)$$

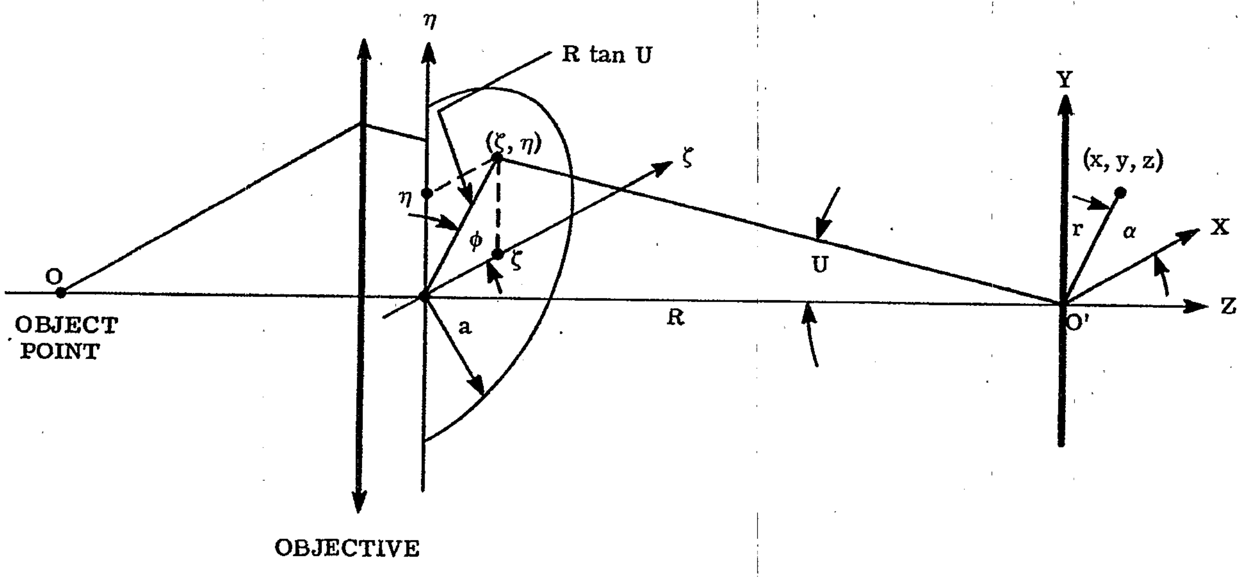


FIGURE 16. 36--Notation with respect to objectives that have axial symmetry and circular apertures of radius a.

Upon substituting from Equations (183), (184), and (185) into Equation (180), one obtains the result

$$F_o(x, y, z) = \int_0^{U_m} \int_0^{2\pi} e^{-ikn\left[\sin U\,(x\cos\phi + y\sin\phi)\right] - z\cos U} \frac{\sin U}{\cos U}\, dU\, d\phi \quad (186)$$

in which U_m is the largest value of U in the cone of axial rays that pass from the object point O to the conjugate point O'. Equation (186) is the Luneburg-Debye statement of the primary diffraction integral.

16.26.1.3 A change of variable from U to the zonal numerical apertures ρ where

$$\rho \equiv \sin U \quad ; \quad \rho_m = \sin U_m \quad ; \qquad (187)$$

renders both the form and the physical interpretation of the primary diffraction integral somewhat simpler. One obtains from Equations (186) and (187) the result

$$F_o(x, y, z) = \int_0^{\rho_m} \int_0^{2\pi} e^{iknz\sqrt{1-\rho^2}}\; e^{-ikn\rho\,(x\cos\phi + y\sin\phi)} \frac{\rho\, d\rho\, d\phi}{1-\rho^2} \qquad (188)$$

Equation (188) is known to hold well for the image space of microscope objectives, telescopes, etc., in which ρ_m is so small that one can set $1 - \rho^2 = 1$ in the denominator. For example, with microscope objectives $\rho_m \approx 3/150 = 0.02$ so that $\rho^2 \lesssim 0.0004$, a quantity that can be ignored in the denominator of (188) but not in the exponential of the numerator except when $z = 0$, i. e., except when one focuses upon the plane which is conjugate to the object point. In computing $F_o(x, y, 0) = F_o(x, y)$ for the conjugate plane $z = 0$, one obtains the Fraunhofer type of diffraction integral

$$F_o(x, y) = \int_0^{\rho_m} \int_0^{2\pi} e^{-ikn\rho\,(x\cos\phi + y\sin\phi)}\; \rho\, d\rho\, d\phi \qquad (189)$$

upon neglecting ρ^2 in the denominator.

16.26.1.4 Typical of diffraction integrals of the Fraunhofer type, the integral (189) is easily integrated. Introduce polar coordinates r, α such that

$$x = r\cos\alpha \quad ; \quad y = r\sin\alpha \quad . \qquad (190)$$

Then from Equation (189)

$$F_o(x, y) = F_o(r) = \int_0^{\rho_m} \int_0^{2\pi} e^{-ikn\rho r\cos(\phi - \alpha)}\; \rho\, d\rho\, d\phi \quad . \qquad (191)$$

As in the integral (159), $F_o(r)$ is independent of α. Furthermore, from Equations (162) and (191)

$$F_o(r) = 2\pi \int_0^{\rho_m} J_o(kn\rho r)\, \rho\, d\rho \quad . \qquad (192)$$

Comparison of Equations (192) and (163) shows that the integral (192) is obtained from the integral (163) by setting $S = 1$ and $a = \rho_m$. Hence, we conclude at once from Equation (157) that

$$F_o(r) = 2\pi\, \rho_m^2\, \frac{J_1(2\pi n\rho_m r/\lambda)}{2\pi n\rho_m r/\lambda} \qquad (193)$$

wherein r is the distance from the diffraction head, and

$$n\rho_m = n\sin U_m \qquad (194)$$

is the zonal numerical aperture of the objective with respect to its image space of refractive index n. We see that $F_o(r)$ is a real number when it is evaluated at the sharply focused image plane $z = 0$ for objectives that have negligible spherical aberration. The time-averaged-energy density $W(r) = |F_o(r)|^2$. Thus,

$$W(r) = 4\pi^2 \rho_m^4 \left[\frac{J_1(2\pi n\rho_m r/\lambda)}{2\pi n\rho_m r/\lambda}\right]^2 \quad . \qquad (195)$$

a result that should be compared with that of Equation (171).

16.26.1.5 The primary diffraction integrals (191), (192), and (193) are called the Airy type, and the corresponding, idealized objectives are distinguished as the Airy type of objective. As in the discussion leading to Equation (170), the first zero of $W(r)$ occurs at $(2\pi n\rho_m r)/\lambda = 3.8317$ or at

$$r = r_1 = \frac{3.8317}{2\pi n\rho_m}\lambda = \frac{0.6098}{n\rho_m}\lambda \quad . \qquad (196)$$

r_1 is the distance from the diffraction head (where $W(r) = W(0) = \pi^2\rho_m^4$, its maximum value) to the first zero of W(r) in the image space. The distance r_1 is frequently utilized as unit distance and is called the Airy unit with respect to the image space. The quantity $n\rho_m = n \sin U_m$ is the numerical aperture of the objective with respect to its image space.

16.27 RESOLUTION WITH CIRCULAR APERTURES

16.27.1 General. It is not possible to specify a universal limit of resolution that applies to all kinds of details in an object field. Resolving power varies with the type of details that are to be resolved, with the manner in which the object is illuminated, with the wavelength utilized for illumination, with the numerical aperture of the objective, and with the degree of correction of the objective. Resolution can depend upon the type of optical system. For example, it can be shown theoretically that an ordinary microscope cannot resolve two nonabsorbing particles, irrespective of their separation, when the optical path difference Δ between the particle and its surround becomes so small that $\sin \Delta$ can be replaced by Δ. The chief reason for this peculiarity is that with such particles, contrast in the image becomes so poor that one cannot actually observe the particles. When the ordinary microscope is replaced by a phase microscope, contrast in the image is increased enormously. Consequently, the phase microscope can exhibit resolving power when the ordinary microscope does not. Finally, resolving power depends upon the criterion that one is willing to accept in concluding from the observation of the image that the details in question are distinct, i. e., are resolved.

16.27.1.1 We shall restrict our considerations of resolving power to the resolution of two self-luminous particles whose dimensions are neglibible. Let one particle be located at point O on the axis as in Figure 16. 37 According to Rayleigh's criterion of resolution, two object points O and P will be resolved provided their separation equals or exceeds the separation r_o for which the maximum energy density in the diffraction image of one particle falls upon the first minimum in the diffraction image of the second particle as illustrated in Table 16. 1. We have seen in the previous section that the distance r_1 from the central maximum to the first minimum is given by Equation (196) for objectives of the Airy type. Hence, the linear limit of resolution is r_1 (or one Airy unit) in the image space where

$$r_1 = \frac{0.6098}{n\rho_m} \lambda . \tag{197}$$

Therefore,

$$r_o = \frac{r_1}{|M|} = \frac{0.6098}{|M|\, n\rho_m} \lambda \tag{198}$$

is the linear limit of resolution in the object space where M denotes the magnification ratio. If the objective obeys the Abbe sine condition,

$$|M|\, n\rho_m = |M|\, n \sin U_m = n_o \sin U_{o,m} \equiv \text{N. A.} \tag{199}$$

where N. A. denotes the numerical aperture of the objective with respect to its object space. Therefore,

$$r_o = \frac{0.6098}{\text{N. A.}} \lambda . \tag{200}$$

The linear limit of resolution, r_o, for two self-luminous object points is one Airy unit with respect to the object space of the objectives that approximate the Airy type.

16.27.1.2 The corresponding angular limit of resolution, θ_1, is given by

$$\theta_1 = \frac{r_1}{V} = \frac{0.6098\, \lambda}{n\rho_m\, V} \tag{201}$$

in which V is the image distance, Figure 16. 37. When U_m is small as in the image space of telescopes and microscope objectives

$$\tan U_m = \frac{D}{2V} \rightarrow \sin U_m = \rho_m . \tag{202}$$

Hence

$$\theta_1 = \frac{(2)\,(0.6098\, \lambda)}{nD} = \frac{1.22\, \lambda}{nD} \tag{203}$$

where D is the diameter of the objective and n is the refractive index of the image space.

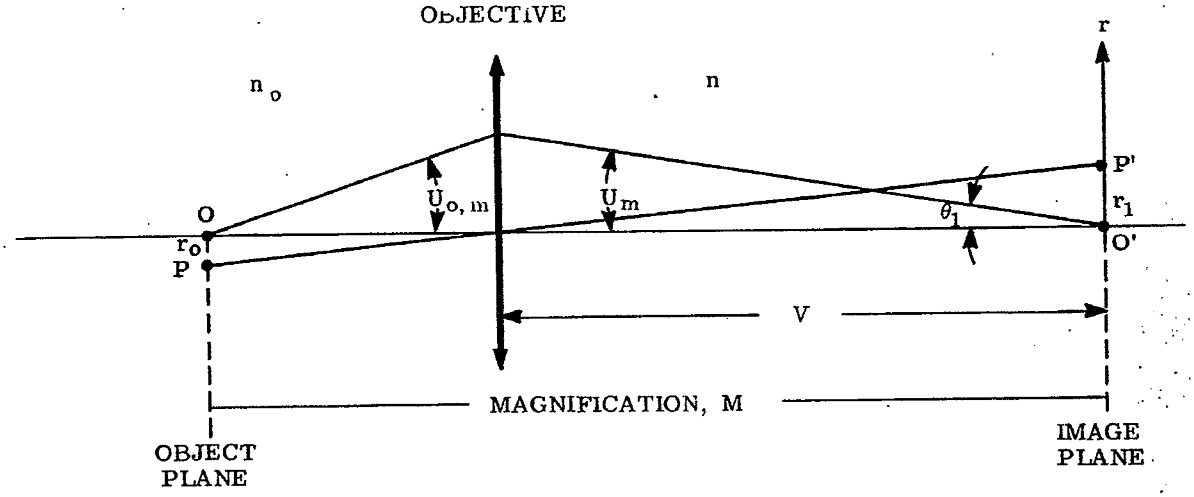


FIGURE 16. 37-Notation with respect to the resolution of two self-luminous object points by objectives having circular apertures.

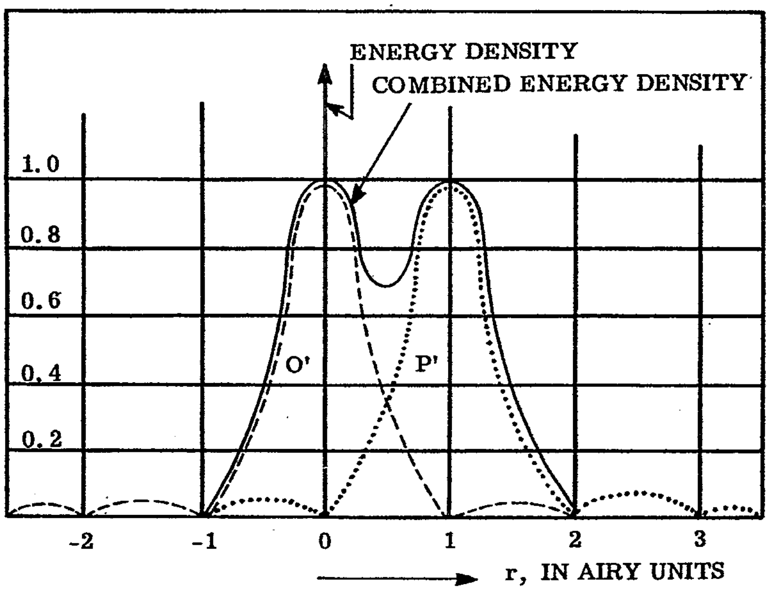


TABLE 16. 1-Physical situation at the limit of resolution based on Rayleigh's criterion. O' and P' are the curves of the energy densities in the image of two, like, self-luminous particles, O and P, respectively. The solid curve is the sum of the energy densities due to the two particles. This solid curve displays an easily seen dip at 0.5 Airy Unit, the mid-point between the geometrical images of the two particles.

16.27.1.3 The limit of resolution obtained from Rayleigh's criterion is a conservative limit with highly corrected objectives. The Sparrow * or physical limit of resolution is 0.78 Airy units for Airy type objectives. In principle, this limit can be approached but not realized. Many observations have indicated that resolutions near 0.81 Airy units have been achieved with highly corrected objectives.

16.28 OUT-OF-FOCUS ABERRATION

16,28.1 General.

16.28.1.1 The out-of-focus aberrations for axial object points are included in Equation (188) in which the system is out-of-focus by the amount z. The integration with respect to $d\phi$ can be carried out just as in the argument leading to the integral (192) even when $z \neq 0$. One obtains instead of (192) the integral

$$F_o(r) = 2\pi \int_0^{\rho_m} e^{iknz\sqrt{1-\rho^2}} J_o(knr\rho)\, \rho\, d\rho \tag{204}$$

when ρ^2 is ignored in the denominator of the primary diffraction integral (188).

16.28.1.2 In the presence of spherical aberration and out-of-focus aberration, one finds in general that $F_o(r)$ is of the form

$$F_o(r) = 2\pi \int_0^{\rho_m} P(\rho)\, J_o(knr\rho)\, \rho\, d\rho \tag{205}$$

for axial object points where $P(\rho)$ is called the pupil function. We see that the pupil function $P_z(\rho)$ corresponding to out-of-focus aberration is

$$P_z(\rho) = e^{iknz\sqrt{1-\rho^2}} \tag{206}$$

Whenever $\rho^2 \ll 1$, it is usual to accept the approximation

$$\sqrt{1-\rho^2} = 1 - \frac{\rho^2}{2} \tag{207}$$

and to write

$$F_o(r) = 2\pi\, e^{iknz} \int_0^{\rho_m} e^{i\pi nz\rho^2/\lambda} J_o(knr\rho)\, \rho\, d\rho \quad , \tag{208}$$

a result that follows from Equations (205), (206), and (207).

16.28.1.3 The following is one of the simplest methods for estimating the maximum tolerable amount z that an objective of given numerical aperture $n\rho_m$ can be out-of-focus. Equation (208) is easily integrated for any axial image point $r = 0$ because $J_o(0) = 1$. Thus

$$\begin{aligned} F_o(0) &= 2\pi\, e^{iknz} \int_0^{\rho_m^2} e^{\frac{i\pi zn\rho^2}{\lambda}}\, \rho\, d\rho \\ &= 2\pi\, e^{iknz} \left[e^{\frac{i\pi nz\rho_m^2}{\lambda}} - 1 \right] \Big/ \frac{i\pi nz}{\lambda} \quad . \end{aligned} \tag{209}$$

The corresponding energy density $W(0) = |F_o(0)|^2$ is now

$$W(0) = 4\pi^2 \left(e^{\frac{i\pi nz\rho_m^2}{\lambda}} - 1 \right) \left(e^{\frac{-i\pi zn\rho_m^2}{\lambda}} - 1 \right) \Big/ \frac{\pi^2 n^2 z^2}{\lambda^2}$$

$$W(0) = 8\pi^2 \left[1 - \cos\left(\pi zn\rho_m^2/\lambda \right) \right] \Big/ \frac{\pi^2 n^2 z^2}{\lambda^2} \quad , \text{ or}$$

$$W(0) = 4\pi^2 \rho_m^4 \left[\frac{\sin\left(\pi zn\rho_m^2/2\lambda \right)}{\pi zn\rho_m^2/2\lambda} \right]^2 \quad , \tag{210}$$

where $W(0)$ is the energy density at the diffraction head when the objective is out-of-focus by the amount z. $n\rho_m$ is the numerical aperture of the objective with respect to its image space.

16.28.1.4 When $z = 0$,

$$W(0) \equiv W_o = 4\pi\rho_m^4 \quad , \tag{211}$$

---

* See H. Osterburg, Microscope Imagary and Interpretation, J. Opt. Soc. Amer,, 40, 299 (1950).

a result that agrees, as it should, with $W(0)$ from Equation (195). Let

$$K = \frac{W(0)}{W_o} = \left[\frac{\sin(\pi z n \rho_m^2 / 2\lambda)}{\pi z n \rho_m^2 / 2\lambda} \right]^2 \tag{212}$$

where K is the ratio of the energy density at the diffraction head when the objective is out-of-focus by the amount z to the energy density at the diffraction head when the objective is in focus. The ratio K is, we note, an even function of z when no spherical aberration is present. The assigned value of K becomes a criterion for the maximum tolerable out-of-focus distance z.

16.28.1.5 Suppose that

$$\pi |z| n \ \rho_m^2 / \lambda \leqq \pi / 2 \ . \tag{213}$$

This means (see Equation (208)) that the phase aberration due to being out of focus shall not exceed one-fourth wavelength. By introducing $(\pi z n \rho_m^2) / 2\lambda = \pi / 4$ into Equation (212), one finds that $K = 0.8106$. Hence, the criterion

$$K \geqq 0.8106 \tag{214}$$

is equivalent to the criterion* of Equation (213). We learn from Equations (213) and (214) that if

$$|z| \leqq \frac{1}{2} \ \frac{n\lambda}{(n\rho_m)^2} \ , \tag{215}$$

the central energy density in the out-of-focus image of a self-luminous object point located upon the axis of the objective will not fall below 81.06 per cent of the maximum central energy density which occurs at the state of sharpest focus $z = 0$. $n\rho_m$ is the numerical aperture of the objective with respect to its image space. Consider, for example, the case in which the refractive index n of the image space is unity and in which $\rho_m \equiv \sin U_m = 0.1$. From Equation (215), $|z| \leqq 0.5\,\lambda/0.01 = 50$ wavelengths.

16.28.1.6 This diffraction theory for out-of-focus images will become less reliable as ρ_m becomes large; but within the range of applicability of the theory, the depth of focus should vary inversely as the square of the numerical aperture of the objective and directly as the wavelength. This conclusion is quite different from that based upon the more elementary notions of geometrical optics.

16.28.1.7 The reader who wishes to examine the applications of the more general primary diffraction integral (205) to cases in which the pupil function $P(\rho)$ includes spherical aberration and in which $r \neq 0$ may consult an excellent, detailed publication** by Guy Lansraux.

---

*This criterion is known as Rayleigh's criterion for phase aberrations.

**Guy Lansraux, "Calcul des Figures de Diffraction des Pupilles de Revolution," Revue D' Optique, 26, 24-45 (1947).

17 OPTICAL MATERIAL

17.1 INTRODUCTION

17.1.1 General. The only properties of an optical material which are used in the actual design of a system are its indices of refraction, and the quantities derived therefrom (such as the ν-value). However, frequently the designer must select materials which meet requirements that are not concerned with the quality of imagery, but which are important with respect to satisfactory fabrication and performance of the instrument. Often he must make a compromise between desirable optical properties, and such characteristics as weight, availability, durability and cost.

17.1.2 Coverage. The following discussion of optical materials will involve the properties of refracting materials, and those of reflecting materials.

17.2 REFRACTING MATERIAL CHARACTERISTICS

17.2.1 Transmission. A refracting material, to be useful, obviously must transmit radiation in the wavelength region in which it is to be used, or it may be used to absorb the undesirable radiation of other wavelengths. In some instances, the refracting material transmits imperfectly in the region of use, and the designer must determine what thicknesses (if any) he can use without greatly impairing the performance of the instrument. The amount of light transmitted through a lens or plate is limited by surface reflection, by absorption within the medium, and by diffusion.

17.2.2 Surface reflection. When light is incident on the boundary surface between two refracting media, part of the light is transmitted into the second medium and part is reflected back into the first. The ratio of the reflected light to the incident light, sometimes called the reflection coefficient or reflectance or

$$R = \frac{1}{2}\left[\frac{\sin^2 (I - I')}{\sin^2 (I + I')} + \frac{\tan^2 (I - I')}{\tan^2 (I + I')}\right], \tag{1}$$

reflectivity, is where I is the angle of incidence and I' is the angle of refraction. If the light is incident normally, so that $I = 0$, and if one medium is air, or a vacuum with index of refraction $= 1$, then the

$$R = \left[\frac{n - 1}{n + 1}\right]^2, \tag{2}$$

expression reduces to wherein n is the index of refraction of the other medium. This formula is the one most frequently used in estimating reflection losses. Since the reflection coefficient depends on the indices of refraction, and these in general depend on the wavelength, the reflection coefficient itself varies with wavelength. Thus for an extra dense flint, having a refractive index of 1.720 at the sodium D line and 1.673 at 2.5 microns, equation (2) yields reflection coefficients of 7.0% and 6.3% respectively at the two wavelengths. Whether the variation with wavelength is important depends on the application. Many times it is satisfactory to use a single value of the index throughout the range of wavelengths under consideration.

17.2.3 Absorption. When radiant energy passes through a medium other than a vacuum, a portion of it is usually converted into another form of energy. This phenomenon is called absorption, and the energy so absorbed is no longer available for image formation.

NOTE

> In this discussion a terminology will be used which is becoming standard. However, the reader is warned that in other publications on this subject the use of terms and symbols may vary considerably with those presented herein. In consulting other published data on optical materials be sure of the exact meaning given to the terms and symbols used, regardless of their apparent similarity to the terms and symbols in this handbook. The basic physical concepts are the same in all cases.

17.2.3.1 According to Bouguer's law, the amount of light not absorbed in the passage through a homogenous medium, i.e., the transmitted light, is a decreasing exponential function of the thickness. That is,

$$W = W_o \; e^{-\alpha t} \tag{3}$$

In this equation W_0 is the original flux. W is the amount remaining unabsorbed after passage through a thickness t of the medium; α is a quantity called the absorption coefficient of the medium. (Note that this equation refers to what happens within the medium - it is not complicated by the reflection losses which occur when the light passes through a boundary surface into or out of the medium.) In general, α is a function of the wavelength.

17.2.3.2 The extinction coefficient κ and the absorption constant k are also commonly used constants. These are related to the absorption coefficient by the equation

$$\kappa = \frac{k}{n} = \frac{\alpha\lambda}{4\pi n} , \tag{4}$$

λ being the wavelength of light and n the index of refraction. For computation purposes it is sometimes convenient to replace the base e in equation (3) by base 10, and write

$$W = W_0 \, 10^{-\beta t} \tag{5}$$

This permits using tables of common logarithms. We shall call β the "absorption coefficient to base 10", when necessary for clarity, and simply the "absorption coefficient" when the distinction is clear from the context. Therefore, $\alpha = 2.303\beta$. The quantity,

$$T = \frac{W}{W_0} = e^{-\alpha t} = 10^{-\beta t} . \tag{6}$$

is called the internal transmittance of the thickness t.

17.2.3.3 The absorption characteristics of a material are usually measured by placing a sample in the form of a polished, plane-parallel plate in the beam of a spectrophotometer, and determining what portion of the radiation incident on the first surface of the sample emerges from the second. This determination is made for as many wavelengths as desired. Most modern spectrophotometers automatically draw a curve of transmittance as a function of wavelength. It is evident that the results from the spectrophotometer include the effects of reflection loss and absorption loss. (They also include the effect of loss due to diffusion, but this is usually negligible in comparison with the others.) It is frequently necessary to be able to separate the two effects in order to predict the benefit to be gained by low-reflection coatings, or the absorption to be expected from other thicknesses. Of the light incident on the first surface of the sample, the fraction R is reflected and lost, and the fraction (1-R) passes into the sample. Of this, the fraction T passes through the sample without being absorbed and reaches the second surface. Here a fraction R is reflected and only (1-R) passes through the surface. That is, of the original radiation, the fraction

$$T_1 = T\,(1-R)^2 \tag{7}$$

passes through the second surface. However, the radiation reflected at the second surface passes back through the medium toward the first surface, where a part of it is reflected back into the medium, and so on. Summing over all such passages one obtains for the total fraction T_∞ passing through the second surface of the sample,

$$T_\infty = \frac{(1-R)^2\,T}{1 - T^2 R^2} . \tag{8}$$

It is the quantity T_∞ (usually expressed as a percent) which is measured by the spectrophotometer.

17.2.3.4 Since R can be determined from the refractive index, the value of T can be computed from equation (8) which can conveniently be rearranged as the quadratic equation:

$$T^2 + \left(\frac{1}{R} - 1\right)^2 \cdot \left(\frac{T}{T_\infty} - \frac{1}{R^2}\right) = 0 \tag{9}$$

If R is not large, the amount of light contributed by the re-passages through the medium may be negligible, and it will then be simpler and satisfactory to use equation (7) for T , using the spectrophotometer reading for the value of T_1. Having determined the value T for the sample and knowing its thickness, one can determine the absorption coefficient from equation (7) or equation (9). Then one may compute the fraction which will be transmitted through other thicknesses, with the same or different surface reflectivities.

17.2.3.5 Consider the following example. Spectrophotometer curves were run on a sample of ordinary plate glass, 6.18mm thick, and on a sample of "waterwhite" plate, 6.6mm thick. At the wavelength of 600mμ the former transmitted 88.3% and the latter, 91.4% (the measured transmittances are significant to a few tenths of a percent). Both glasses have an index of refraction of about 1.52 at this wavelength. How would their transmittances compare in thicknesses of 25.4mm, if the surfaces in use are to have low-reflection coatings with reflectivities of 1% ? For the ordinary plate glass

$$R = \left(\frac{0.52}{2.52}\right)^2 = 0.043 \quad .$$

Substituting in equation (7) yields

$$T = 0.964.$$

Use of the more precise equation (9) would give $T = 0.960$ but the difference is hardly significant in view of the limited precision of the initial data. Then,

$$\beta = \frac{-\log\ 0.964}{6.18\text{mm}} = 0.00257/\text{mm} \quad .$$

The internal transmittance of a 25.4mm thick piece would then be

$$\text{antilog}\ (-0.00257 \times 25.4) = 0.860 \quad .$$

Since the coated surfaces are to have reflectivities of 1%, substitution in equation (7) shows that 84% of the incident 600mμ radiation would pass through the plate. For the waterwhite plate the surface reflectivity is again 0.043. Substitution in equation (7) yields

$$T = 0.998 \ .$$

This value is so near unity that the difference is not reliable in view of the limited accuracy of the data from which it was computed. However, it is safe to assume that T_1 is no worse than 0.995 and that hence

$$\beta \leq \frac{-\log\ 0.995}{6.61\text{mm}} = 0.00033/\text{mm}.$$

Consequently the internal transmittance of the 25.4mm thickness will be at least 0.981 and with the surfaces coated at least 96% of the radiation will pass through the piece.

17.2.4 Diffusion. Some light, in passing through a medium is deviated from its path due to the presence of fine inhomogeneities in the medium. This effect is called diffusion. In extreme cases it causes the medium to be translucent rather than transparent. Some of the diffused light is lost from the optical system. That which remains is not image-forming but, being spread over the image area, reduces contrast. For this discussion it is not necessary to consider the physics of the phenomenon beyond remarking that the amount of scattering is a function of the ratio of the size of the inhomogeneity to the wavelength of light, the amount decreasing as the ratio decreases. (The effect can be noticed in a long line of automobile headlights at night, unless the air is very clear. Since more light is scattered from the blue end of the spectrum, the distant lights look more yellow or orange than the near ones.) As a result, inhomogeneities which would be bothersome in the visible region may be of negligible importance in infrared work.

17.3 REFRACTIVITY AND DISPERSION

17.3.1 Selection of materials.

17.3.1.1 From the available media which transmit satisfactorily in the wavelength region with which he is concerned, the designer must select those with index and dispersion characteristics best suited for his needs. In the visible region, it is usually sufficient for the designer to know the indices of refraction for a few conventionally specified wavelengths, and to do much of his calculations with the quantities derived from them, (such as the ν value and the partial dispersion ratios). Glass-makers' catalogs customarily describe the refractive properties of the glasses in terms of these standard quantities.

17.3.1.2 In the ultraviolet and infrared regions, procedures and requirements are not so well standardized. Use of standard wavelengths for index measurement, and of dispersion constants based on such measurements, has not become common. It is often necessary to work from such tables of values of refractive index as are available, and to interpolate for the wavelengths one wishes to consider in the design. To aid in selecting either the derivative $dn/d\lambda$ or the related quality

$$\frac{(n-1)}{\frac{dn}{d\lambda}} .$$

The latter is analogous to the ν-value; for achromatism of a thin doublet, the ratio of the powers of the two elements should be the negative of the ratio of the values of $(n-1)/(dn/d\lambda)$ of the two media. Use of the former is similar to the use in the visible region of $n_F - n_C$; the ratio of the total curvatures of the two elements of a thin achromatized doublet should equal the negative reciprocal of the ratio of the two derivatives. Some publications tabulate values of the derivative $dn/d\lambda$.

17.3.2 Optical homogeneity.

17.3.2.1 It is important that the refractive index of a lens or prism be constant throughout the piece. Usually the requirement for uniformity within the piece is more rigid than the requirement for uniformity from piece to piece. There are two principal causes of optical inhomogeneity: chemical inhomogeneity and improper annealing.

17.3.2.2 Since most glassy substances are complex mixtures, rather than precise chemical compounds, it is difficult to make them chemically homogeneous throughout. The presence of streaks of slightly varying composition results in striae. The harm done by striae depends on their loaction in a system. If they are so placed and oriented that all parts of each image-forming ray bundle pass through about the same optical path in the striae, the effect may be negligible. Thus a moderate amount of flat striae, approximately parallel to the plane of a weak lens, may be tolerable. On the other hand, in a reflecting prism in which the beams pass through the same volume in at least two different directions, it is impossible to meet this condition, and material for such prisms should be free of striae.

17.3.3 Mechanical strain. The thermal history of the piece of material may also cause optical inhomogeneity. If it has been such as to result in the presence of mechanical strains, the material becomes locally polarizing. The presence of this defect can be observed by examining the piece between crossed polarizers. However, even if there is no detectable strain, the material still may be optically inhomogeneous, and a more careful fine annealing of the glass, to accomplish the effect known as compacting, is necessary. (The fine annealing is also necessary to bring the index of the glass to its maximum value. Melt sheets supplied by manufacturers generally give index and dispersion measurements made on fine-annealed samples). Heating the glass to softening for making molded blanks, slumpings, etc. cancels out its previous thermal history and necessitates re-annealing to ensure the quality desired.

17.3.4 Optical isotropy and anisotropy.

17.3.4.1 In most cases the designer requires that his materials be optically isotropic - that is, the index of refraction at any point must be constant, regardless of the direction in which the radiation is passing the point. Only occasionally does he want anisotropic or polarizing media. Optical glasses are isotropic, except as made locally anisotropic by improper annealing. Many crystalline materials with otherwise attractive characteristics are anisotropic and hence unsuitable for making lenses and non-polarizing prisms. However, crystals belonging to the cubic system, when free from mechanical strain, are optically isotropic, and some such are noted below.

17.3.4.2 Occasionally a weakly anisotropic material such as sapphire can be used satisfactorily, by orienting its optic axis parallel to the optic axis of the system.

17.3.4.3 A crystal which is isotropic optically is not necessarily so in all mechanical properties, and it may be desirable to have blanks for lens making cut with preferred orientation with respect to the cleavage planes. This will ensure uniform grinding or minimize losses from fracture along the cleavage planes.

17.4 INCLUSIONS

17.4.1 Imperfections in optical materials. Optical materials may have imperfections in the form of tiny opaque or refracting inclusions. In ordinary optical glass the most common of these are called, according to their nature, bubbles, seeds or stones. Other refracting media may have similar imperfections. The harm done by such defects depends on their position in the optical system. An inclusion near an image plane, as in a field lens or in the plane-parallel plate on which a reticle is formed, will appear as a bothersome out-of-focus object in a visual system, or may give a false signal in an infrared system. The same inclusion in an objective lens may have negligible effect on the performance. Tolerances on such imperfections should be set with the specific use of the part in mind.

17.5 ENVIRONMENTAL CHARACTERISTICS

17.5.1 Optical system requirements. Withing recent years, requirements for the performance of optical systems under extreme environmental conditions, such as in airborned equipment or under exposure to desert, jungle and artic conditions, and also the necessity of using available refracting materials other than ordinary optical glass outside the visible spectrum, have made it necessary for the designer to be conscious of the thermal, mechanical and chemical characteristics of the materials he proposes to use. The most important of these are the following:

(1) Softening characteristics - cold flow. A lens or window obviously should hold its shape, including the figure of its refracting surfaces, under storage and service conditions. The softening temperature of the material should be high enough to ensure that this will be the case. A few materials exhibit the phenomenon of cold flow, a tendency to deform even at ordinary temperatures.

(2) Resistance to thermal shock. Materials vary in their ability to undergo rapid changes in temperature without fracture. Some media, unless precautions are taken, are likely to crack during ordinary grinding and polishing. Others can withstand the changes involved in exposure on the exterior of supersonic airframes. The larger the piece of material, the more subject it is to damage from this cause.

(3) Coefficient of thermal expansion. This characteristic, while related to the ability of a material to withstand thermal shock, is also important if the instrument must withstand a wide range of temperatures. The coefficient material should be matched with that of the cell in which it is to be held or to which it is to be cemented.

(4) Specific gravity. Especially in airborne equipment, weight is an important factor. Hence knowledge of the specific gravity of the material is useful. However, the significance of the specific gravity depends on the effect of the optical characteristics of the design of the system. If, for example, a material of higher density permits using a single lens instead of two, there still may be a weight advantage with the dense material.

(5) Hardness. The hardness of the material is important both during fabrication and in service. A very hard material, such as fuzed quartz, is difficult and time-consuming to grind. On the other hand, a very soft material is likely to develop scratches or sleeks during polishing. Soft materials should be avoided in locations where they will be exposed to surface abrasion, as in exposed domes in airborne equipment.

(6) Surface deterioration. Polished optical surfaces may deteriorate from a number of causes, the susceptibility depending on the material. Staining due to the action of atmospheric moisture and carbon diozide is known as weathering. Closely related to weathering is susceptibility to tarnish or etching by weak acids, which may frequently occur in the atmospheres encountered. In hot, humid climates mold may grow on the surface, leaving marks which cannot be removed except by re-polishing.

(7) Devitrification. Some glassy materials have a tendency to devitrify, and the tiny crystals formed make the glass diffusing. Glasses for instruments which must be stored for long periods should be chosen with this in mind.

17.6 REFRACTIVE MATERIALS FOR SPECIFIC WAVELENGTH RANGES

17.6.1 Classifications. It is convenient to discuss materials under the following classifications:

(a) for the visible region including the near ultraviolet and the near infrared, from about 0.36μ to about 2.20μ.

(b) for the ultraviolet region at wavelengths shorter than 0.36μ.

(c) for the infrared at wavelengths longer than 2.2μ.

17.6.2 Applicable materials for visible spectrum. Most optical design work is done in the visible spectrum and more types of media have been developed for it than for the others. These may be classified as ordinary optical glasses (including rare earth glasses), crystals (natural or synthetic), and plastics. Some of the media used mainly for the other wavelength ranges are transparent in this region and could be used here, but their disadvantages make them unattractive as compared with optical glass.

17.6.2.1 Optical glass. The properties of the ordinary optical glasses are well catalogued, and the designer should obtain the catalogs of several manufacturers for reference. These lists vary widely in the amount of information provided. The most elaborate lists gives indices of refraction for a number of wavelengths distributed through the visible region, the Abbe or ν-value, and several partial dispersions and dispersion ratios, along with information on specific gravity, weathering characteristics, amount of internal imperfections, and thermal characteristics. There is much similarity between the glasses of various manufacturers. However, if a catalog does not give the desired information on a characteristic which is critical for a special application, it is well to inquire of the manufacturer rather than to take data from another manufacturer's catalog.

17.6.2.1.1 Figure 17.1 shows the range of index values and ν-values within which most of the commercially available glasses fall. However, manufacturers' lists vary widely in the variety offered. The available varieties are adequate for most purposes. If the importance of the project warrants the expense, melts of glasses with properties intermediate between those listed can sometimes be arranged for.

17.6.2.1.2 As pointed out elsewhere in this handbook, complete chromatic correction of a simple system puts a condition on the partial dispersion ratios of the glasses involved, as well as on the ν-values. Thus, in a doublet, to bring light of three wavelengths to a single focus, the ν values of the two glasses should differ, but the partial dispersion ratios should be equal. Unfortunately the partial dispersion ratio for most glasses (and other substances as well) is practically a function of the ν-value, so this requirement cannot be met. However, a few manufacturers offer a small number of glasses which depart from the rule sufficiently to be useful.

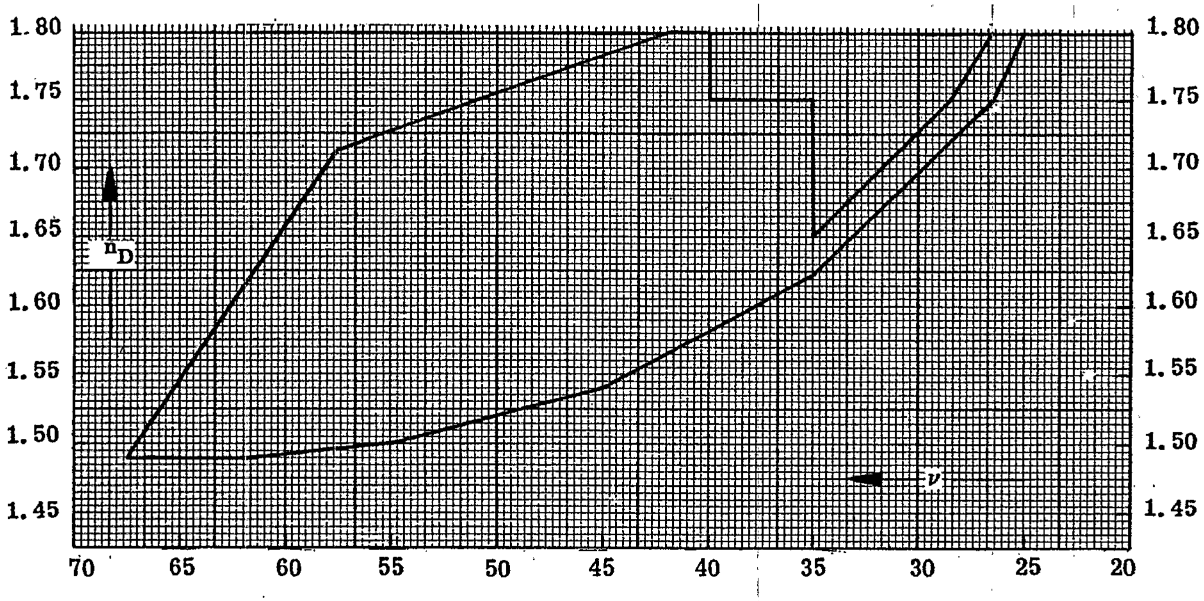


Figure 17.1- The range of commercially available glass, with respect to index and ν values.

17.6.2.1.3 Although one foreign manufacturer's catalog has recently begun listing index values for the wavelengths 0.365μ in the ultraviolet and 1.014μ in the near infrared, most manufacturers give only values in the visible region and to 0.405μ in the ultraviolet. In selecting glass types for the ultraviolet or the infrared without resorting to actual measurement it is necessary to resort to some sort of extrapolation. Kingslake and Conrady* measured the indices of 17 glasses at various wavelengths from 0.365μ to 2.5μ. An approximate value of the index of another glass can usually be derived by interpolation between the values for glasses with similar n_D and ν in Kingslake and Conrady's list. A number of formulae for demonstrating the refractive index as a function of wavelength have been proposed, the constants of the formula being determined from the indices of the glass in the visible region. A fairly recent formula has been proposed by Herzberger.**

17.6.2.2 Crystals. Crystals, natural or synthetic, are rarely used for this wavelength region. The principal application is in the reduction, or elimination, of secondary spectrum in apochromatic systems. Fluorite (CaF) has been used for many years for this purpose since it has an index $n_D = 1.434$, and a $\nu = 95.4$. The partial dispersion ratio

$$\tilde{P} = \frac{n_D - n_C}{n_F - n_C} = 0.293$$

which is equal to that of an ordinary glass with $\nu = 57$ to 59. Synthetic crystals, with diameters up to 7.5 inches are available. Synthetic alum crystals ($KAl(SO_4)_2 \cdot 12H_2O$) has also been used. For this material $n = 1.456$, $\nu = 58.2$, $\tilde{P} = 0.315$. Since alum is highly water soluble, the material cannot be used in exposed locations. It has usually been made the central member of a cemented triplet.

17.6.2.3 Plastics. In recent years there has been considerable interest in some of the synthetic resins which can be made transparent and colorless. These materials have a number of attractive features, foremost of which are its lightweight, and the possibility of fabricating elaborate forms quite inexpensively by casting.

17.6.2.3.1 Specific gravities of typical optical plastics are 1.05 to 1.19, compared with 2.48 for crown glass and 3.41 for dense flint. The weight savings are obvious.

17.6.2.3.2 Since the optical elements can be fabricated by casting or molding (processes which are not practical with glass for any, except crude optical elements) it is possible to aspherize mold inserts in order to aspherize lenses, and to cast optical elements complete with mounting flanges, or other mechanically convenient additions.

17.6.2.3.3 Counterbalancing these advantages, available plastics have a number of disadvantages which so far have worked against their use in any but very simple and non-critical systems. They are quite soft, and scratch easily. Many attempts have been made to improve this defect, such as coating the lenses with harder materials, but low scratch resistance remains an important defect. They are quite subject to change of index with humidity, and to significant change of index and surface figure with temperature. They also deform easily under mechanical force, and some tend to turn yellow with age.

17.6.2.3.4 The principal injection-molding materials now used are acrylics, with $n_D = 1.49$, $\nu = 58$, and styrenes with $n_D = 1.59$, and $\nu = 31$. Allyl-diglycol carbonate is used for casting.

17.6.2.3.5 Cast spectacle lenses have some popularity because they are light weight, but the principle defect preventing their wider acceptance is that they scratch easily.

17.6.2.3.6 Large cast blocks are used in unit-power, tank periscopes to shorten the optical path through the periscope, but the end prisms, which ideally should be cast in a single piece with the block, are still made of glass.

17.6.2.3.7 Acrylic lenses have been used in large numbers in simple one and two lens slide viewers, and in inexpensive camera viewfinders. Their quality is adequate for these applications, and their lightness and cheapness are attractive. An additional advantage in the view finders is the aspherization of one lens surface to obtain the elimination of barrel distortion. Injection-molded acrylics are also widely used as singlet meniscus objective lenses in box cameras. One camera manufacturer has introduced a plastic f/8 triplet. It is possible that more applications of this sort will be made as fabrication processes are developed and improved.

---

* R. Kingslake and H. G. Conrady, J. Opt. Soc. Am. 27; 257 (1937).
** M. Herzberger, J. Opt. Am. 32; 70 (1942).

17.6.3 Materials suitable for wavelengths longer than 2.2μ.

17.6.3.1 For application reasons, optical systems in the infrared have been designed mainly for three wavelength regions: the region near 1μ, for use with image converter tubes and infrared photography; the region from 2 to 3μ, for use with lead sulphide cells; and the region near 4.2μ. (Some applications have called for coverage of longer ranges.) Current activity shows an interest in longer wavelengths. As noted above, ordinary optical glasses are satisfactory for the region near 1μ. However, they begin to absorb strongly at about 2.5μ, and their usefulness in the 2 to 3 micron region depends on the requirements of the application, and on the thicknesses needed. For longer wavelengths, other materials must be used.

17.6.3.2 The search for satisfactory infrared transmitting materials has been active and is continuing vigorously. The situation is complicated by the fact that many of the applications, for airborne equipment and especially for windows for such equipment, require excellent mechanical characteristics, and large pieces of material.

17.6.3.3 The properties of approximately fifty materials which are of potential usefulness in the infrared, and which constitute nearly all such materials which had been investigated up to the end of 1958, have been gathered and tabulated by Ballard et al* and the designer should provide himself with a copy of this reference. It lists for each material, to the extent to which information was available at the time of publication, the composition, molecular or atomic weight, specific gravity, crystal class, transmission, reflection loss, refractive index, dispersion, dielectric constant, melting temperature, thermal conductivity, thermal expansion, specific heat, hardness, solubility and elastic moduli. The transmission is usually presented as a curve showing external transmittance as a function of wavelength. Refractive index as a function of wavelength is given in tabular form. The dispersion (which is, of course, implicitly contained in the refractive index table) is for many substances plotted as a curve showing the derivative of index with respect to wavelength as a function of wavelength. One chapter is devoted to tables each listing the substances arranged in order with respect to a single characteristic such as thermal conductivity or coefficient of linear expansion, thus permitting easy comparison. The last chapter is devoted to a brief discussion of glasses and plastics.

17.6.4 Materials suitable for wavelengths shorter than 0.36μ. Work in ultraviolet optics is much less active than in infrared, and the existing applications for the most part impose much less stringent requirements on non-optical characteristics of the materials, and in size of pieces required. A modest number of suitable materials is available, some of them synthetic crystals. Important ones are listed in Table 17.1 together with some of their properties. Index and dispersion in the visible region are given to show the general optical position of the material. Literature references to ultraviolet index and transmission information are included.

Material	N_D	ν	Cutoff	Max. Piece Diameter	Remarks
Sodium Chloride	1.544	42.8	0.25μ	7.5 in.	Highly water soluble
Potassium Bromide	1.560	33.4	0.21μ	7.5 in.	"
Potassium Iodide	1.667	23.2	0.25μ	7.5 in.	"
Lithium Fluoride	1.392	99.3	0.11μ	6.0 in.	
Calcium Fluoride	1.434	95.1	0.125μ	6.0 in.	
Fused Quartz	1.458	67.8	0.22μ	several in.	
Barium Fluoride	1.474	81.8	0.145μ		

Table 17.1 - Materials suitable for ultraviolet beyond 0.36 μ.

17.7 REFLECTING MATERIALS

17.7.1 Thin films. Nearly all reflecting surfaces in optical instrumentation are made by forming thin films, usually by evaporation but sometimes by chemical means, on glass or some other appropriate substrate. Most frequently simple metal films are used. For special purposes, multilayer films are sometimes provided, which give enhanced reflectance in a particular wavelength region, or may serve as filters, reflecting in one region and transmitting in others.

---

* Stanley S. Ballard, Kathryn A. McCarthy, William L. Wolfe: Optical Materials for Infrared Instrumentation; IRIA Report #2389-11-S, L'. of Michigan, Ann Arbor, 1959.

17.7.2 First surface versus second surface coatings. Depending on the application, the radiation may be incident on the surface either on the side in contact with the glass, or in the side exposed to the air on a vacuum. The latter use is commonly called a "first-surface" reflection; the former, "second surface." When used as a second surface reflector the film can be protected by such means as plating and painting. For use as a first surface reflector, any protecting coating must be transparent. A film of silicon monoxide is frequently employed. Such a film is so thin as to have very little effect on the reflectance in the visible and in the infrared out to 9 or 10µ. However, the thickness required for protection of the surface is sufficiently large to produce interference effects which may decrease the reflectance at various ultraviolet wavelengths.

17.7.3 Simple metal coatings. The situation with respect to simple metal coatings has been summarized by Haas*. Aluminum, silver, gold, copper, and rhodium are considered to be the most important mirror metals. The only material that has a high reflectance in all useful regions, the ultraviolet, visible and infrared, is aluminum. The reflectance of all other metals drops rapidly in the visible or ultraviolet. In the near infrared between 1 and 2µ the average reflectance of silver, gold, copper, and aluminum is higher than 95% but the reflectance of aluminum is about 2% to 3% lower than that of the other three materials. In the far infrared at 10µ all four metals have a reflectance of 98% to 99% and even rhodium reflects about 96%. Today, the most frequently used high reflecting coating for first surface mirrors is vacuum deposited aluminum. It adheres better to glass and other substrates than the other high reflecting materials, it does not tarnish in normal air, and it is very easy to evaporate. Obviously, aluminum coatings are especially important for astronomical mirrors and reflection gratings where high reflectance in the ultraviolet is required. Figure 17.2 of Haas (loc cit) shows the reflectance of freshly deposited films of Ag, Al, Au, Cu and Rh as functions of the wavelength from 0.22 to 10µ. For second-surface reflectors in the visible and near infrared silver is frequently used. As a second-surface reflector it can be adequately protected by copper-plating and painting.

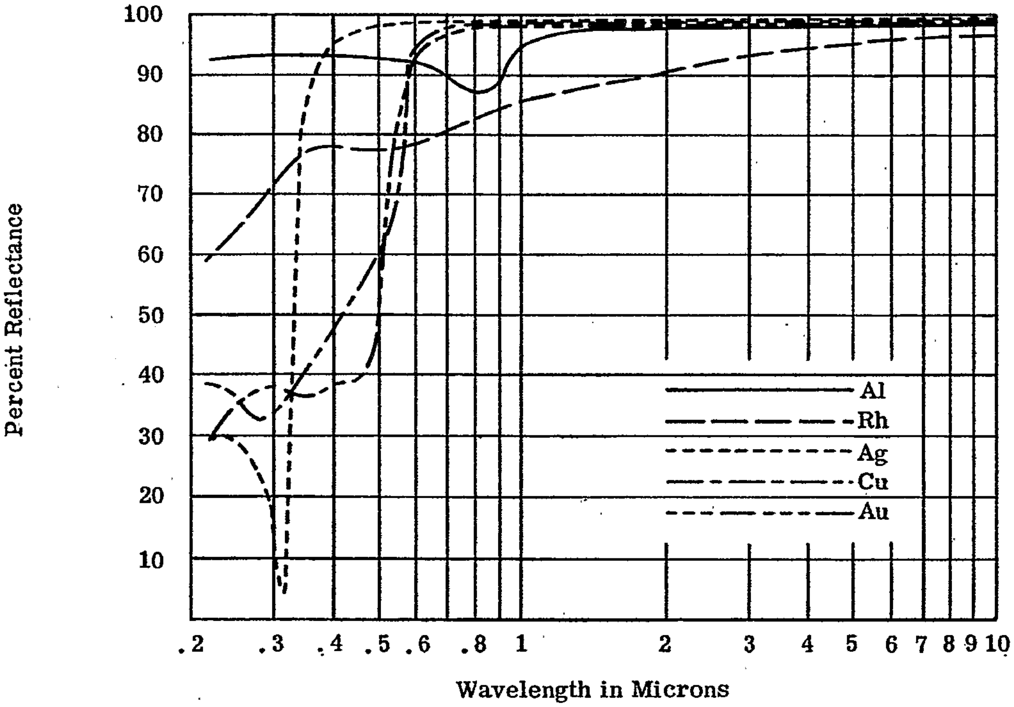


Figure 17.2- Reflectance of freshly deposited films of Ag, Al, Au, Cu, and Rh as function of wavelength from 0.22 to 10µ. (From Jour. Optical Soc. America, G. Haas 945; 945, 1955)

---

* Haas, George: Filmed Surfaces for Reflecting Optics; JOS 945-945, November, 1955.

17.7.4 Reference. The subject of reflective and anti-reflective coatings is treated extensively in Sections 20 and 21 of this handbook.

17.8 AVAILABILITY; COST; EASE OF WORKING

17.8.1 General. Materials with attractive optical properties are sometimes unsuited to particular applications because of unavailability, either in quantity or in size of pieces required, because of high cost, or because of the difficulty of working the material satisfactorily. Even in cases in which the cost itself is not an objection, the designer of military instruments should as far as possible avoid the use of materials which would be in critical supply, or cause excessive demands on manpower in times of national emergency.

18 ATMOSPHERIC OPTICS

18.1 INTRODUCTION

The use of optical instruments, and one may include the eye in this definition for the present purpose, is limited by the transmission of the atmosphere. As will be developed in the following section, the degree to which the usefulness of an instrument is limited by the atmosphere is roughly related to the aperture. Under specialized conditions however, aperture is of no consequence since image information content is reduced to zero by scattering.

Two generalized types of optical systems may be considered with regard to atmospheric effects. First is the information gathering type of system which depends upon image formation for fulfillment of its purpose. Second are photometric devices which have as their purpose only the detection of amount of radiant energy. Except for highly specialized instrumentation, most optical systems in which atmospheric contributions are significant are of the first class. This discussion will deal only with instruments intended for the former purpose.

18.2 EXTINCTION

18.2.1 Types of transparency losses contributing to extinction.

18.2.1.1 The light of the stars, planets and sun is observed to be weakened as it penetrates the earth's atmosphere. The investigation of this effect, termed the Astronomical Extinction, is one method which has been used in investigating the scattering properties of the atmosphere. Further information has been obtained by measuring the distribution of light in the daylight sky. Both of these types of measurement are of course repeated under laboratory conditions with artificial light sources, and are then useful in the study of extinction produced by media such as rain and fog. The extinction in the clear sky has three contributing components:

(1) Absorption by air molecules.

(2) Rayleigh scattering by air molecules.

(3) Scattering by dust and haze.

The scattered component observed from the clear sky is due to factors 2 and 3 only while the first factor includes the continuous Chappuis bands of Ozone that cover a substantial section of the visual spectrum. Also included in the first factor are the molecular absorption bands of Oxygen, Water Vapor, Carbon Dioxide etc. These later bands are of particular interest to those dealing with Infra-Red optical systems. The so-called atmospheric "windows" are merely areas without such absorption. The design of Infra-Red systems is considered a highly specialized application, and further discussion of this type of atmospheric absorption will therefore not be included here. Atmospheric transmission as a function of wavelength is presented in the International Critical Tables.

18.2.1.2 Extinction can be measured by observing the intensity I of a light source as seen through a volume with scattering particles. If I_o is the intensity of the same light source seen through the same volume without scattering particles, the ratio is:

$$I/I_o = e^{-l\kappa} \qquad (1)$$

where l is the path length through the particulate medium and κ is the extinction coefficient. Even with a very dilute smoke, for example, a considerable intensity of scattered light I_s may easily be detected at right angles to the direction of transmission. A certain amount of true absorption does of course occur, and this represents the actual disappearance of light -- the energy of which is lost as heat. (The kinetic energy is transformed into heat motion of the molecules of the absorbing material.) The term κ may therefore be considered as consisting of two components, a (absorption) and s (scattering). The lateral scattering of a beam of light as it passes through a cloud or aerosol may be easily demonstrated. This phenomenon is closely associated with both diffraction and reflection. The light scattered at right angles to a primary beam is found to be partially plane polarized. The reason that the scattered component normal to the primary beam is not completely polarized is because of multiple scattering where light is scattered several times and therefore yields a non-constant plane of polarization. The integrated effect is therefore one of partial polarization. The polarization produced in the scattered component has been used as the basis of a compass designed by Pfund. The fact that the maximum polarization is observed perpendicular to the incident sunlight allows the approximate direction of the sun to be measured even though the sun itself be invisible. This information coupled with time data allows direction to be computed in areas where magnetic data are unreliable.

18.2.1.3 The transmissivity of the atmosphere for information-gathering optical systems is an inverse function of the extinction and is limited by particulate scattering such as that caused by haze, clouds and fog, and dusts. The subsequent discussion covers the extinction and, more particularly, scattering by these media.

18.2.2 Particulate scattering.

18.2.2.1 Two of the three factors which contribute to the extinction in the clear sky are concerned only with scattering:

(1) Rayleigh scattering by air molecules.

(2) Scattering by dust and haze.

In the unclear or turbid sky, additional scattering is produced by rain, clouds or fog. If one disregards absorption by air molecules, the extinction may be regarded as a function of the above two listed causes.

18.2.2.2 The first quantitative evaluation of the laws governing the scattering of light by small particles was made by Lord Rayleigh in 1871. His mathematical investigation yielded a general law for the intensity of scattered light. The law derived is generally applicable to particles of any index of refraction different from that of the dispersing medium. The one restriction on the application of the law is that the particle size must be greatly smaller than the wavelength of light.

18.2.2.3 As might be almost intuitively arrived at, the intensity of the scattering is found to be directly proportional to the incident intensity and also directly proportional to the square of the volume of the scattering particle that is typical of the scattering medium so long as the maximum particle dimension is small with respect to a wavelength.

18.2.2.4 A most interesting result of the work of Rayleigh is that the degree of scattering is dependent upon wavelength. Thus, for a given size of particle, long waves are scattered less than short ones, because the particles present obstructions to the light waves which are smaller compared to the longer wavelengths than to the shorter ones. The general expression is given as:

$$I_s = k\lambda^{-4} \tag{2}$$

where k is a constant of proportionality and I_s the intensity of the scattered component. For example, red light at a wavelength of 0.72 microns has a wavelength 1.8 times as great as that of violet light at 0.40 microns. The law predicts:

$$I_{s_v} = (1.8)^4 \quad I_{s_r} = 10 \quad I_{s_r}$$

assuming that the particles doing the scattering are small compared to either wavelength. As was pointed out previously, if the particulate scattering is due to large particles, the wavelength dependence does not follow the law expressed in equation (2). The relative intensity of the scattered component (I_s) as a function of wavelength is shown in Figure 18.1.

18.2.2.5 It is because of this wavelength dependence that so-called "haze" filters are used. The scattered blue light in the sky may be removed by use of a minus-blue filter so that the sky will appear darker in photographs. Indeed, a dark red filter -- corresponding to the wavelength of least scattering -- will show the clear sky as nearly black. Much here of course depends upon the definition of "clear" sky. When the particulate size is such that it is no longer small with respect to the wavelengths involved, white light scattering will occur. This is the result of ordinary diffuse reflection from the surface of the particles. When transparent large particles are considered, more complex results are obtained. In general however, the final result is that white light is scattered as white light and not to a greater extent at the shorter wavelengths.

18.2.2.6 Haze, together with dusts, forms the atmospheric contaminant referred to as the "aerosol". The aerosol consists of airborne particles of microscopic and sub-microscopic size. This aerosol contributes to scattering and therefore to the extinction.

18.2.2.7 The scattering pattern for atmospheric haze cannot be arrived at analytically with the same accuracy as the extinction law given previously. This is due to the fact that secondary scattering occurs. Therefore, the skylight cannot be simply separated into a factor due to molecular scattering and one due to haze.

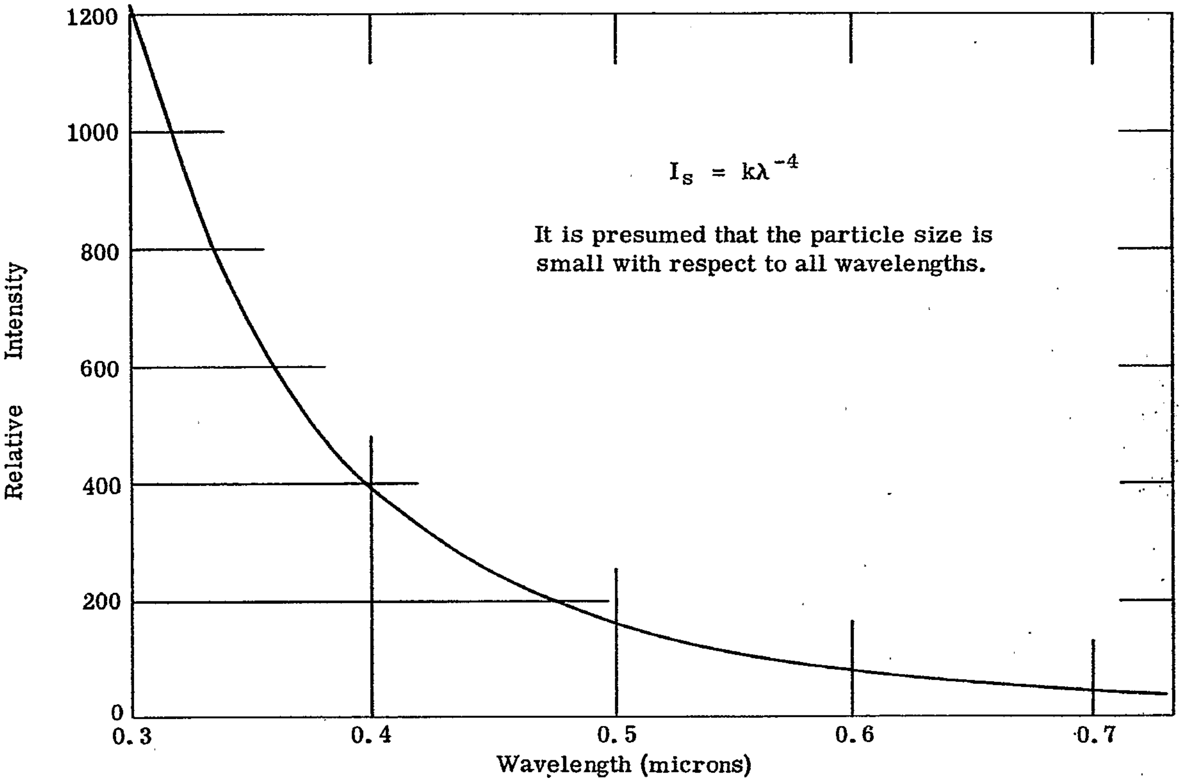


Figure 18.1- Relative intensity vs wavelength.

18.2.2.8 There are two special sources which contribute to the world-wide aerosol. These are volcanic eruptions and forest fires. The eruption of Krakatoa in 1883 was the cause of beautiful sunsets in most parts of the world. A large eruption at Iceland in 1783 caused a particularly red sun to be observed. A more recent eruption of Katmai in 1912 was of special interest since it allowed measurements of particle size to be made. The particles of haze were from 1.0 to 1.2 microns in diameter. Dust particles of the size 0.34 to 0.44 microns diameter were also found and measured.

18.2.2.9 Summer heat haze is a special case. It covers large portions of the earth's surface which are free from dust, fires, or even human habitation. For example, a more or less dense haze hangs over jungle areas in Colombia and the Amazon basin the year around. In the summer months, the heat haze covers most of the continental United States except the desert areas of the Southwest. This haze is of such a density that the open country is almost as hazy as the cities. This heat haze has nothing to do with so-called smog, fog, smoke or even moisture. It is just as dense near the ground, where the relative humidity may be 40 per cent, as near the inversion layer 10,000 feet up, where the humidity may be near 100 per cent. This blue summer haze looks so much like the man-made "Los Angeles" smog that it may almost be considered a "natural" smog. The most usual explanation given for this phenomenon is one based on the production of the "natural" smog by organic emanations from plants. These plants release vast quantities of material into the air. Most of the aromatic substances emitted by plants are hydrocarbons, many of them essential oils. Also, many plants contain terpenes, which after evaporation produce the pungent odors of the deserts and pine forests. Since such summer heat haze is not found over water areas, this explanation is logical.

18.2.2.10 Consider the blue light of a clear sky as seen at sunset. Observing directly overhead, the light scattered will be partially plane-polarized with a maximum perpendicular to the incident sunlight. The reason that completely polarized light is not observed is due to multiple scattering -- scattered light that is scattered a second, third time etc. before reaching the observer. The observation of a red sunset is attributed to the scattering of light by fine dust and smoke particles in the earth's atmosphere near the surface. Since the amount of atmosphere traversed by the sunlight is much greater at sunset due to the low angle, the dust path to the observer is greatly increased with the result that blue and violet have been scattered and the sun appears in the remaining colors. A by-product of the sunset effect is that, due to the low total illumination reaching an observer from the sun and the relatively large amount still present from sky light, the spectral shift in illumination is toward that quality of light yielded by the sky. Under these circumstances, the sky overhead is still blue. The shift is therefore toward the blue end of the spectrum. Due to the so-called "blue myopia" various vision difficulties may be encountered at this time of day even though the total illumination level is quite adequate. The optical instrument designer must consider the use of a spectral balancing filter in visual instruments to be used at dusk. The obvious problem of total light level vs. detectability requires compromise which must be decided upon the merits of each particular application.

18.2.2.11 The drops of rain, clouds, and fog are much larger than the particles that go to make up dusts and haze. A cloud or fog may contain a proportion of very small droplets, but the diameters of the drops that form the largest portion of the cloud or fog are those that dominate the extinction characteristics of the medium. These particle diameters range from 10 to 40 microns.

18.2.2.12 One extremely important fact to be drawn from this, as may be concluded from the mathematical discussion, is that the extinction contribution of clouds, fog and rain where particle sizes are from 10 to 40 microns or larger is virtually constant throughout the ultraviolet, visual, and near infrared regions. Exactly where one may establish the maximum of the extinction curve will fall is difficult to predict. Measurements throughout the 0.4 to 5.5 micron range have been made by many observers. The extinction remains constant within ± 10 per cent. The efficacy of special filters, for example, under these conditions is not real if penetration is considered. A more complex case exists where cloud cover is not complete, and it is desired only to obtain the best contrast condition against this broken cover. In "typical" rain, the droplet size is of course again much larger than in the case of fogs and clouds. Again, the extinction is essentially constant throughout the spectrum. It is for this reason that special color filters are of little or no value in the penetration of clouds, rain or fogs wherein the particle size is of the order of from 10 to 40 microns. A slight gain is obtained by the use of filters in visual instruments which peak the response to fit the response curve of the eye. Likewise, the response of a given photographic material or photo-cathode substance may be matched in order to give some small penetration advantage. Filters are of a definite advantage whenever the particle size is small with respect to the wavelength.

18.3 EXTINCTION AND VISUAL INSTRUMENTS

18.3.1. Imposed limitations on visual instruments.

18.3.1.1 The main limitation produced by extinction exists in two degrees. First, if the transparency is very low (extinction coefficient high) no art exists whereby penetration in the visual spectrum may be achieved. This is the case in dense fogs, clouds and rain. Since it is impossible to design for this case, the best compromise must be made for the second condition, namely; when transparency is reduced by smaller concentrations of aerosol or water droplet dispersions. We will consider instruments of limited aperture and power. Such instruments will not be seriously hampered by poor "seeing" conditions (such as will be discussed subsequently), but will be limited by extinction.

18.3.1.2 Visual optical systems having apertures of up to two inches and magnifications up to 10 diameters are not greatly affected by differences in air density, inhomogeneity, turbulence etc. when used at elevation angles in excess of 15 degrees. Larger systems are more seriously affected, nearly proportional to their aperture.

18.3.2 Filters.

18.3.2.1 The main thing which can be done in visual optical instrument design to optimize performance under less than ideal atmospheric conditions is the addition of a series of suitable filters. The purpose of the instrument will of course govern the choice of filter combinations made available.

18.3.2.2 The color filter which will prove most advantageous depends upon the object being viewed and upon the background. For ground objects, where the background and subject are both subjected to the same scattering effects, filter choice will depend upon the color of the object. For light or dark objects against the sky, the filter choice will depend upon the brightness of the target and the color of the sky.

(1) When the target is white and bright against the clear blue sky, a filter (red or yellow) which will make the sky appear relatively darker with respect to the object is desired.

(2) Observation of dark objects against a blue sky presents a more complex problem. Some observers maintain that if the blue myopia of the observer is compensated, best results are obtained with a blue filter. Others, and the writer is one of them, fail to observe any improvement in visibility of dark objects against the blue sky with any filter, and subsequently suggest that a "clear" condition is best under these circumstances. The same is true of the compromise required in the observation of mixed white and dark objects against the clear blue sky. For visual purposes the "clear" condition is best or nearly so.

(3) While a white target is usually lighter than a clear blue sky background, it may easily be darker than a white sky background. In this instance it is found that color filters are of little or no help. If the target is colored, the choice of filter to obtain the best result is that which darkens the target more than the background.

18.3.2.3 As we have noted, a great proportion of scattered light is partially polarized perpendicular to the incident radiation causing the scattering. The use of polarizing filters may allow an increase in contrast under special conditions where the scattered component may thusly be at least partially eliminated.

18.3.2.4 At the risk of redundancy, it will be pointed out once again that when large particle media are producing scattering, no filter will appreciably increase penetration. In the case of true haze, where the particle size is small, removal of the scattered light is practical by filtration. A minus-blue filter, or even a red filter may be used under these circumstances to improve clarity.

18.3.3 Light gathering power.

18.3.3.1 One chief result of high background illumination is reduced contrast when bright objects are to be viewed. Also, intervening particles cause scattering which further reduces contrast. The later cause of contrast reduction applies to both bright and dark objects viewed against a comparatively bright background. If the visual consequences of this loss of contrast are to be minimized, it is necessary that further contrast loss not be produced by having an exit pupil smaller than the pupil of the eye. It is found to be advantageous, therefore, to accomplish effective aperture reduction by means of neutral density filters rather than by reduction in aperture by means of adjustable or fixed aperture stops. The use of such neutral density filters has a further advantage, not directly connected with atmospheric optics, in that the resolving power of the instrument is not reduced as it might well be if the aperture were reduced by physical limitation of diameter.

18.4 EXTINCTION AND PHOTOGRAPHIC INSTRUMENTS

18.4.1 Effect of instrument orientation. Photographic instruments, for the present purpose, will be placed in two groups -- those used looking up, such as missile tracking cameras, ballistic cameras etc, and those looking down, such as aerial cameras, satellite borne cameras etc. The atmospheric effects due to extinction coefficient variations are quite different in the two cases. Some generalizations can be made which apply to both groups, but the two are better treated separately.

18.4.2 Photographic instruments "looking up".

18.4.2.1 Missile and satellite tracking cameras and telescopes always have the sky as a background. Depending upon the conditions of a particular firing or satellite pass, the sky may be either clear or cloudy, bright or dark. Due to a gain in performance at greater scale, missile tracking and satellite surveillance cameras are usually of great focal length. Satellite acquisition cameras, on the other hand, are of high optical speed and short focal length.

18.4.2.2 As would be expected, the contrast between the target and the sky background is a function of the distance from the camera to the target. Horizontal and vertical components of this distance however, are not of equal importance in the reduction of contrast. For a dark target on a clear day (blue sky background) subject contrast is decreased more by an increase in altitude than by an increase in horizontal range. For a bright target on a clear day the opposite is true.

18.4.2.3 As is the case for visual instruments, the greatest gain in performance is obtained through the use of suitable filters. The type of filter to be used with an optical system "looking up" depends upon the relative contrast of the object and the background.

(1) Filters for photography of white targets against a blue clear sky require effective darkening of the sky background and heightening of the relative brightness of the target. For a spectrally non-selective white target, the use of a red filter (Wratten 25 or 29) ordinarily gives excellent contrast results. A blue filter will give poor results in this situation since this will effectively reduce contrast. A yellow filter will yield satisfactory results under these conditions, but the red filter will be superior provided that the filter factor or effective density is not so great as to reduce the available illumination below operable limits.

(2) Filters for photography of dark targets against a blue clear sky require effective darkening of the target and heightening of the relative brightness of the background. This can be achieved by the use of a blue filter (Wratten 47). The photographic material does not suffer from the blue myopia of the visual observer. Also, photographic contrast in blue is essentially the same as for other colors -- in fact for certain color materials more tone scales are available from the blue sensitive layer than from any other -- so that the contrast rendition can be excellent under these circumstances if a blue filter is used. If the target is in an orientation 180 degrees opposed to the sun, and if the sun is very low on the horizon, a yellow or red filter may actually be better than the blue, due to the blue deficiency of the sunlight which must pass through a greatly extended atmospheric path. In the case of a satellite that is effectively outside the earth's atmosphere however the choice of filters is more complex, since the scattering path length for light reaching the optical system from the target is nearly the same as the scattering path length of the light from the sun (here the atmospheric path length only is considered).

(3) When photography is done on either white or gray objects against a sky background which is whitened by scattering, filters are found to be of little if any assistance. If the target is colored, the choice of filters should be such as to insure the darkening of the target more than the background. For example, a blue filter would be used with yellow or red targets.

18.4.3 Photographic instruments "looking down".

18.4.3.1 It is well known that photographs taken with large and small lenses of similar quality have nearly the same microscopic contrast in the presence of haze, but have ground resolution modules approximately proportional to focal length. There are many other factors to be considered, such as the quality of the lens, the lens mounting, the type and speed of the shutter, the filter used if any, etc. In general however, if all these factors are equal or nearly so, and if the laboratory performance resolution wise (in lines per millimeter) remains independent of focal length as is usually the case with present day aerial lenses, then the focal length is the most important factor in obtaining a given desired ground resolution.

18.4.3.2 The only extinction or scattering phenomenon which can be partially eliminated by means of filters in the case of photographic instruments "looking down" is haze. Minus-blue filters are commonly employed for this purpose. Since most ground objects have extremely low contrast ratios (of the order of log 0.1 and thereabouts) the elimination of haze effects is extremely important. Red filters have a better haze elimination characteristic in the case of true haze, but these have a tendency to cause natural greenery to appear too black for satisfactory interpretation. The usual compromise has been to select a yellow or orange filter for the purpose of haze minification.

18.5 "SEEING"

18.5.1 Atmospheric factors affecting "seeing".

18.5.1.1 "Seeing," as differentiated from transparency, (reciprocal extinction) is concerned with those factors which limit the information content of images by causing a lack of bulk homogeneity in the optical medium preceding the optical system.

18.5.1.2 At least two basic causes for differences in refractive index of air may be demonstrated. First, there is the "suryp" analogy, in which the density -- and thus the refractive index -- is changed by local differences in temperature throughout the air mass. A second method of producing differences in refractive index is by the condensation and rarification of air by sonic means. When a sound wave moves through air, the air mass instantaneously consists of a series of sections of air having in turn increased and decreased densities -- and thus increased and decreased refractive index. These two causes of inhomogeneity in air are the basic causes of "seeing" difficulties.

18.5.1.3 Any observer who has made observations with a medium size or larger telescope is aware that the performance of his instrument is seriously limited by the astronomical "seeing". Images of stars are much larger than is the case if the image diameter were to be limited by the diffraction of the telescope objective alone. Lunar and planetary detail is badly smeared when the seeing is poor. For example, the average seeing disc of the Hale telescope of 200 inch diameter is about 2.5 seconds of arc while theory indicates that the resolution should be on the order of 0.04 seconds of arc. The optical quality of the telescope is not at fault. The difference is due to the quality of the "seeing." And, it must be remembered that the location of the 200 inch telescope was chosen for the good "seeing" conditions existing on Mt. Palomar.

18.5.1.4 The total amount of light received from a bright star by a telescope of moderate size fluctuates in an irregular fashion, and in a manner which can be shown to be due not to any intrinsic fluctuations in brightness of the star, but to the fact that the starlight must pass through the atmosphere of the earth wherein there are regions of density irregularity. A 12 inch aperture telescope, for example, will exhibit variations in intensity of ± 10 per cent of the average value. The frequency of variation may change from several seconds per cycle up to thousands of cycles per second.

18.5.1.5 Two types of effect are attributable to differences in the air mass preceeding the objective. The first consists of oscillation or image motion. This is due to the movement of relatively large air masses at comparatively low velocities. The second is scintillation. Scintillation is the fluctuation in the light of stars known to be caused by motion across the telescope objective of a complex pattern of light and dark bands known as shadow bands or the shadow pattern.

18.5.2 Oscillation.

18.5.2.1 The change in position of an image in the focal plane of an objective system due to atmospheric disturbance is known as oscillation. This image defect does not of necessity destroy visual resolution. There is a good likelyhood that photographic resolution will be seriously curtailed if substantial amounts of oscillation are present however, since the photographic plate is inherently an integrating device and does not by itself compensate for shifts in position of a high quality image.

18.5.2.2 On a simplified basis, oscillation may be thought of as being caused by the passage of various lens shaped or prism shaped "chunks" or modules of atmosphere in front of the objective. If each air module is of a size greater than the diameter of the telescope objective, a comparatively good image will be seen whenever the entire objective diameter is covered by a single homogeneous module. Since such air modules occur in various layers at different altitudes and move with different velocities, it is obvious that seeing becomes a very complex thing that is difficult of analysis.

18.5.2.3 In a real situation, several cross-currents of air may contain air modules of various sizes. When, in accordance with the laws of chance, the optical path through a diameter covering the objective is equal to within the tolerance of one quarter wavelength of light, an excellent image will be found in the focal plane of the telescope.

18.5.2.4 In the case of relatively slow moving air masses which give rise to oscillation, some sort of compensation is practical. Photometric guided tracking instruments have been developed to compensate for oscillation. These instruments detect the angular movement of the image of an astronomical object, and move the photographic plate or the image so that the effective position of the image on the plate is constant, and a good photographic image is produced.

18.5.2.5 The larger the telescope, the less the probability that the air mass over it will be homogeneous within a quarter wavelength path difference at any given time. Thus follows the hard fact that smaller telescopes may frequently give better resolution than larger ones even though their theoretical resolving power is not nearly so great.

18.5.2.6 Even when the air modules passing over the objective are of sufficient size so that theoretical or nearly theoretical resolution may be obtained, the image motion caused by the shifting air masses is such that long period photography may fail to yield anything approaching what may be seen visually. Very short exposure photography has been tried in efforts to circumvent this difficulty. Also, guidance of the photographic plate and/or the image forming beam have been tried to yield better photographic results. Both of these methods have yielded some success. Excellent planetary photographs have been taken recently with the 60 inch aperture reflector on Mount Wilson after the adaptation of a seeing compensator at the photographic focus. This device moves the final image in accordance with the image shift so that a non-moving image on the plate is the result. Short exposures using telescopes of large aperture have also given promising results in planetary photographs.

18.5.3 Scintillation.

18.5.3.1 Those fluctuations in the light of a star that are not due to any inherent change in brightness of the object but instead to the motion across the telescope objective of a complex pattern of lights and darks, known as the shadow bands or shadow pattern are called scintillation. This shadow pattern is caused by the passage of starlight through atmospheric irregularities which must occur at a considerable height. These irregularities diffract the light and cause rarification and reinforcement of the wavefront at various points along the ground. A fairly complete theory of the relationships between the atmospheric irregularities and the pattern of lights and darks produced in the shadow pattern has been developed.

18.5.3.2 It is common knowledge that the theoretical diffraction pattern can only be observed with telescopes of small aperture and under good conditions of "seeing." With instruments of moderate size -- 36 inches and upwards -- such theoretical resolution is rarely if ever achieved. Using the 36 inch aperture example, the theoretical resolution limit would be less than 0.15 seconds of arc. In practice, however, the starlight is usually spread over a disk of from 2.0 to 5.0 seconds of arc in diameter. This is called the "seeing disk." This "seeing disk" is simply the circle of confusion for the rays reaching the focus (physical optics disregarded for the moment), and its diameter is a measure of the lack of parallelism of the rays when they arrive at the objective from the star. We may thus choose to consider the "seeing disk" as the summation of the diffraction patterns formed by each element of the objective and the air column over it. These elementary diffraction patterns are in rapid oscillatory movement both along and at right angles to the optical axis of the telescope. If the aperture of the large telescope is stopped down to the aperture which would yield the theoretical limit of resolution equal to the actual resolution of the large system, the usual diffraction rings will become clearly visible and sharply defined. The amplitude of brightness scintillation will be noted to have increased roughly inversely to the ratio of new to old diameters. This leads to the conclusion that the "seeing disk" formation is a phenomenon largely independent of brightness scintillation. It seems probable that the "seeing disk" arises from refraction of the rays in their passage through our atmosphere, while brightness scintillation is mainly a diffraction effect arising from the presence of much smaller reinforcement and rarification irregularities within the beam.

18.6 THERMAL EFFECTS

18.6.1 Types.

18.6.1.1 Differences in refractive index due to fluctuations in density which are in turn due to thermal effects play a considerable role in the limitation of vision and photography through optical instruments used for ground level observations. This so-called "ground seeing" frequently limits what can be seen even with low power instruments. In some instances these effects seriously reduce the information which can be gathered by such a small aperture optical system as the human eye.

18.6.1.2 Also, thermal effects in and around larger optical systems are frequently the limiting factor in the performance of these systems. Thus, tube currents and dome currents in astronomical and missile tracking telescopes may reduce the performance of the instrument by a factor of two or more if measures are not taken to circumvent the degradation.

18.6.1.3 The "mirage" or atmospheric striae noticed over the desert floor is an example of what may occur when an air mass is heated by radiation and conduction. In desert areas, even low power telescopic systems of small aperture are very limited in use at low elevation angles. A pair of 7 x 50 binoculars will show much image degradation due to this heat shimmer.

18.6.2 Tube currents.

18.6.2.1 Insofar as the final image is concerned, it matters little whether the density discontinuity occurs without or within the tube of the optical instrument itself. Any telescope tube exposed to thermal radiation or temperature differences of any kind will have variations in density of the air within itself. When these differences become large enough, air flow will occur between the dense and rare areas setting up tube currents. Typically, the air just inside the outside covering of the tube is heated or cooled most rapidly, and a laminar convection current forms wherein air circulates from the perifery to the center. If the tube is an open one, the warmed air -- considering that the tube is being heated by the surroundings -- passes out the end of the tube. Provided the tube is sufficiently larger than the free aperture of the optical system, the air which flows out the end will disturb that remaining in the tube but little. This argument has been presented in favor of open tubes. Unfortunately the issue is not nearly so clear cut. The (assumed) warm air flowing out from the opening of a tube will, since it must obey the gas laws, rise after exiting from the tube. Unfortunately, the path in which it must rise is directly in front of the direction of pointing of the instrument, and while the warm air from the top of the tube causes no difficulty in this regard, that from the bottom flows directly past that area which it is desired to protect from any density differences.

18.6.2.2 A solution which answers the objections to both the closed and open tube arrangements is found in the evacuation of the light path volume. The degree of vacuum need not be high in order to accomplish a substantial increase in performance. A striking experiment may be readily performed to illustrate this fact. If a relatively small telescope of three or four inch aperture is set up for knife-edge test by autocollimation, nearly complete degradation of the image forming qualities of the instrument may be produced by heating the tube of the instrument with a small flame. If the tube is then evacuated to a pressure of 1 PSIA the image quality will be restored to near that present before the heat was applied. It is assumed that the objective is sufficiently thick as to withstand the pressure differential without optical deformation or that a thick optical window has been placed near the objective.

18.6.2.3 The minimization of thermal heating of any tube is desirable from a tube deformation and image quality standpoint. It is good design practice to require some air space beyond the actual optical clearance lines in most optical instruments. A layer of insulation inside the tube also tends to even out the thermal effects so that, while heating will still occur, the unevenness of the heating will not augment thermal disturbances.

18.6.2.4 Most of the energy reaching the earth from the sun is contained within the spectral region between 0.36 and 2.0 microns. The use of highly reflective paints is desirable when thermal effects from radiation are to be reduced. The once prevalent idea that solar heating came mostly from the infrared and that therefore only good infrared reflectivity was required of an instrument paint is quite wrong. The total energy (insolation) from the sun may be as high as 0.028 calories per square centimeter per second. Over 90 per cent of this energy is in the wavelength region below two microns, and nearly 80 per cent is in the region below one micron. So, while good infrared reflectivity is desirable in an instrument white, good visual reflectivity is certainly equally required.

18.6.3 Dome currents.

18.6.3.1 As is the case with telescope tubes, observatory and other housing domes have currents associated with their construction and situation. Particularly in the case of domes opened in the daytime for use, currents may be of such magnitude as to render the housed instrument nearly useless.

18.6.3.2 Painting and insulation are commonly resorted to minimize the effects of thermal heating of the dome. Again, a highly reflective white paint with a reflectivity which matches the solar spectrum as well as possible is desired. Insulation produces a thermal lag which, unless coupled with temperature control as noted later, may actually impair performance rather than improve it.

18.6.3.3 Dome currents may be nearly eliminated if the air inside of the dome is maintained at the same temperature as the air outside. This is an isothermal situation not unlike that in the isothermal jacket of an accurate calorimeter. Since, under conditions of solar radiation, the interior of the dome will usually be considerably warmer than the outside air, it is necessary to provide refrigeration if the desired state of isothermal conditions is to be obtained. At times however -- such as sunrise -- the interior temperature will be much below that of the outside, so that heating capability is also required. It is not sufficient to grossly heat or cool the air within the dome. It is essential that the mixing of air be complete and that inhomogeneities do not exist in the air mass within the dome itself.

18.6.3.4 Next to air temperature control, the best practice is to allow the dome air to arrive at some approximation of equilibrium with the outside air before the housed instrument is to be used. Before observations begin, astronomers open the dome for some period of time to allow the dome, telescope and associated equipment to come to a steady thermal state.

18.7 ATMOSPHERIC CONTAMINANTS

18.7.1 Sources and effects.

18.7.1.1 In addition to the natural products which go up to make the aerosol -- water, dust, smoke etc. -- there are man-made atmospheric contaminants which may produce very undesirable effects on seeing and transparency. For example, industrial smog may limit aerial photography. Certainly the sky light from large cities has greatly reduced the effectiveness of the observatories located near them. Here the man-made effect is two fold. Industrial smog scatters the light which is a by-product of the city so that both transparency and contrast are reduced in that area.

18.7.1.2 Atmospheric contamination by radioactive particles and dusts dispersed by explosions can be viewed as a potential source of seeing degradation if the quantity ever reaches sufficient levels to produce light scattering and even perhaps a very low level of direct radiation. As things now stand, astronomers are counting single photons in the course of measurements on some stars. In these circumstances it is obvious that background light must be kept at an absolute minimum.

18.7.1.3 The increasing amount of carbon dioxide in the atmosphere has undoubtedly increased the "greenhouse" effect present in the atmosphere. While this can have no noticeable effect on visible observations, the increase in infrared absorption may cause detectable variations in the performance of infrared systems. Also, if the increase is sufficient, local heating in areas of high carbon dioxide concentration may cause seeing deterioration due to thermal gradients.

18.8 EFFECT OF ATMOSPHERIC OPTICS ON INSTRUMENT DESIGN

(1) A knowledge of atmospheric optics is important to the optical instrument designer so that he may take advantage of suitable filters, paints and housings in the overall instrument design.

(2) The designer who knows that seeing conditions limit the performance of a telescope more than does the theoretical resolving power will not specify aperture on the basis of theoretical resolution without first considering the actual conditions of use of the instrument.

(3) The design of an optical instrument must go beyond the physical boundaries of the device, it must include at least an approximation of the air column which forms just as much a part of the instrument as does the objective lens. When the limitations placed on performance of high quality optical systems by atmospheric optics are considered in the design phase, a more satisfactory instrument cannot help but result.

19. OPTICS FOR MISSILE TRACKING

19.1 INTRODUCTION

19.1.1 Functions. The main function of optical missile tracking instrumentation is to determine, with precision, the location and trajectory of missiles and satellites. A second important function is to observe the physical appearance and orientation of the space object, and its alterations over short periods of time. In addition, the instrumentation must provide a permanent record of the object's flight for study and later analysis of the trajectory.

19.1.2 Problems.

19.1.2.1 As can be seen from the preceeding paragraph, the problems of tracking and recording the object's flight are closely related to those in astronomy, particularly to those encountered in the observation of planets and planetary detail. However, in missile tracking, additional problems are encountered, since the objects to be observed are not precisely known with respect to location and trajectory. In the proper solution to the problems involved, many contradictory requirements exist, and the correct choice must be made among these requirements for a design necessary to fulfill any particular specification. These requirements are listed as follows:

1. The field of view must be wide enough so that the missile image is picked up.
2. The relative aperture must be high enough to see and record the image under the prevailing lighting and atmospheric conditions.
3. The physical aperture must be large enough so that the Rayleigh limit does not apply to the detail desired.
4. The focal length must be long enough to have sufficient detail appear in the final image.
5. The recorded images must properly follow one another fast enough to determine the trajectory, and/or to disclose the physical condition of the missile.

19.1.2.2 A few remarks are in order concerning these points. The requirement for wide field (1) is so contradictory to all the rest that separate instrumentation in the form of tracking telescopes are needed for picking up the object. In almost all cases, two tracking telescopes are used, as shown in Figure 19.1, one for azimuth and one for elevation and each with its own operator. The function of these operators is to keep the

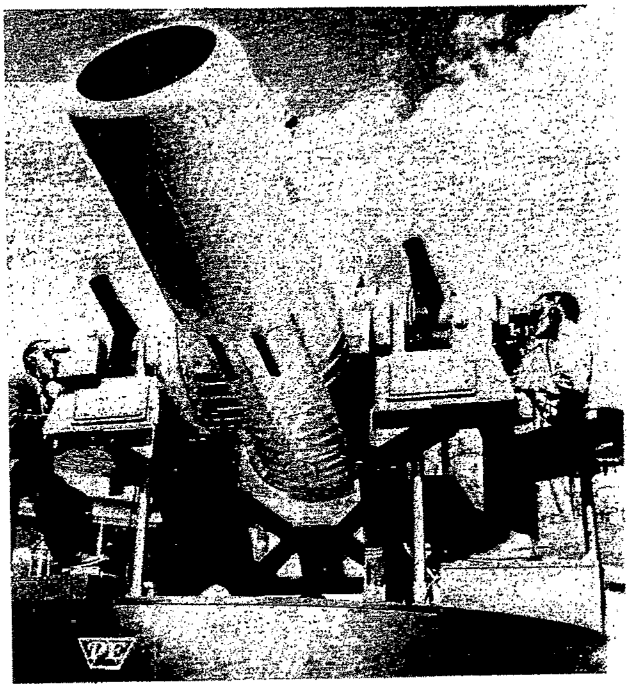


Figure 19.1- Tracking telescopes used with ROT1.
(Courtesy of Perkin Elmer Corp.)

space object close enough to the cross lines so that it will be within the much narrower field angle of the main optical system. This main optical system is completely photographic in nature, so that the final information is recorded on film. Also recorded on the same film, are certain necessary data, particularly an accurate time signal, so that the data from at least two separated stations may be correlated to determine the trajectory of the object by means of triangulation. If errors are to be minimized, a third station may be employed, and suitable data reduction and evaluation means, by high speed computers, used to obtain the trajectory of the object to a high degree of accuracy.

19.1.3 Scope. In this section, the main optical photographic system, which may be either refractive or reflective, will be the basis of the discussion. In addition, emphasis will also be given to those particular optical devices which have been used to obtain desirable properties. The relative merits of refractive and reflective systems will not be gone into here, except to say that for large aperture systems, the arguments are all in favor of the reflective system. In the region of smaller physical apertures, i.e., up to 12 inches diameter, the advantages of greater stability, greater depth of focus, and superior diffraction pattern of the refractive system may overcome the much lower size, weight and freedom from color aberrations of the reflective systems. At larger apertures, however, the balance is so decidedly reversed that the reflective systems have enjoyed an almost complete monopoly in the first half of the twentieth century in the field of astronomy. More advanced designs are now being proposed and constructed in which combined reflective and refractive systems will be of increasing importance, as will be seen from some of the systems described herein.

19.2 REFRACTIVE SYSTEMS

19.2.1 Correction of secondary aberrations. The problems of designing purely refractive systems, of high performance over a relatively small field angle, has been treated extensively in Section 1 of this volume, as has the necessity of correcting for primary spherical aberration, coma, and axial color. Off-axis aberrations, such as astigmatism, distortion, and lateral color, are usually of less importance for a narrow field angle, although they must not be entirely neglected. However, when the focal lengths become large, i.e., in excess of 12 inches, and the relative apertures are of the value of f/8 or lower, great consideration must be given to the correction of the so-called secondary aberrations, the most important of which are secondary spectrum, zonal spherical aberration, and sphero-chromatism.

19.2.2 Secondary spectrum. The methods of handling this troublesome aberration have been discussed in Section 11. For the relatively large apertures we are concerned with in this discussion, the three-glass method leads to fairly high curvatures and impractically thick crown elements. This is also true in combining ordinary crown glass with the crown-flints, where the difference in the Abbe constants is relatively small. It is hoped that glass will eventually be manufactured which will help to alleviate this troublesome aberration. Some years ago, the writer of this section was able to achieve a fairly decent paper design, for secondary color correction, by combining dense barium crown glass with one of the Eastman glasses (EK-320) which gave a higher Abbe spread. However, success was frustrated by the inability to obtain the latter glass, at that time the only available of its type, in better than low-grade C quality. Presently, however, manufacturers are claiming better quality for their lanthanum equivalents, and the situation may change. In addition, other designers are now finding that the later KZF glasses of Schott show considerable promise in this regard and it is imperative therefore, that the designer keep abreast of the newest glass types.

19.2.3 Use of the air space. We turn now to a discussion of the two remaining secondary aberrations namely, zonal spherical aberration and sphero-chromatism, together with a qualitative explanation as to how these aberrations may be minimized or eliminated. With the very long focal lengths involved in the missile tracking problem, either of these can be large enough to ruin the image on the axis. Therefore, the solution to the control of these aberrations is through the judicious use of the air spaces in the interior of the optical design. While these aberrations are of particularly high importance in the design of missile optics, the same considerations apply for other applications throughout the entire field of optics, and a thorough discussion of the corrective properties of the air space is very much in order.

19.2.4 Zonal spherical aberration.

19.2.4.1 In order to understand how the air space may be used to correct for zonal aberration, a short discussion on the nature and origin of the zonal bulge is in order. The spherical aberration of a single surface is given by the expression

$$\mathrm{Sph} = a_1 y^2 + a_2 y^4 + a_3 y^6 + \dots\dots\dots, \qquad (1)$$

where y is the height of the ray above the axis at the surface, and the a's are constant and all of the same sign. The first term is the primary or "Seidel" term and the remaining ones are of so-called higher order.

19.2.4.2 If we now plot the spherical aberration of a single refracting surface, it will look like the solid line in Figure 19.2. The dotted line is the Seidel contribution and since the latter is restricted to the first term in equation (1) it will be a pure parabola. The interval between the solid and dotted lines is due to the presence of higher order terms.

19.2.4.3 Now consider how the zonal bulge originates in so simple an optical design as the corrected doublet. Such a doublet (crown leading) is shown in Figure 19.3. Surfaces 1 and 3 are undercorrecting and relatively weak in curvature as compared to surface 2 which is overcorrecting. It is to be expected that 1 and 3 surfaces will have largely Seidel contributions, while the contribution of the second surface will be relatively rich in higher order terms. If we bend the doublet so that the marginal aberration is zero, we show the sum of the spherical aberration of the undercorrecting surfaces 1 and 3 in Figure 19.4 (a), and that of the overcorrecting surface 2 in (b). The requirement for corrected marginal aberration dictates that $OM_1 = OM_2$. However, the preponderance of Seidel aberration in Figure 19.4 (a) results in a near parabola for the curve, while the presence of higher orders in Figure 19.4 (b) gives a more extreme type of curve as shown, with a milder parabola (dotted extension) plus higher orders. Adding abscissas such as along the line XX to give the final result shown in Figure 19.4 (c) reveals a spherical aberration curve with characteristic zonal bulge, with a maximum of undercorrection at approximately 0.7 of full aperture.

19.2.4.4 Figure 19.5 shows a single positive lens with a marginal ray M and a zonal ray Z. This lens has a large amount of spherical undercorrection and thus the M ray crosses the axis at F_m and the Z ray at F_z. The two rays intersect at the point P. If we set the entrance height of the Z ray at 70 percent of that of the M ray, this ratio is very nearly maintained when the rays leave the lens. However, this percentage ratio gradually increases towards the right and finally reaches 100 percent at the point P, actually reversing beyond this point. If we consider this lens as the first of the doublet shown in Figure 19.3 the negative element of the doublet is located so that the zonal ray Z traverses it at very nearly the original 70 percent height. However, if we allow an air space to exist between the elements as shown in Figure 19.6, the negative overcorrecting element is located at a position where the Z ray is at a height in excess of the original 70 percent value as related to the marginal ray. It thus is acted upon by the negative lens at a point higher than its original assigned value in the aperture, and receives a trifle more overcorrection than it would have received were it a height of 70 percent of the marginal ray. This overcorrection of the zonal ray can be adjusted to reduce or completely eliminate the zonal bulge, or even reverse it, if desired. This control of the zonal bulge obviously depends upon the amount of undercorrected spherical aberration produced by the positive lens and of

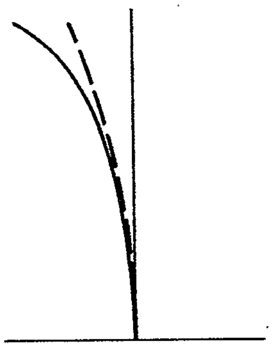


Figure 19.2- A plot of spherical aberration of a single refracting surface.

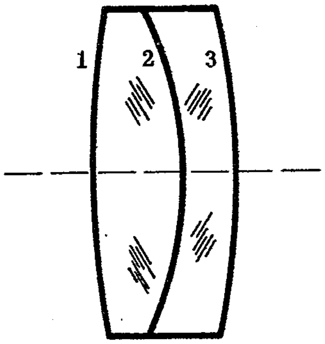


Figure 19.3- Corrected doublet-surfaces 1 and 3 undercorrecting, surface 2 overcorrecting.

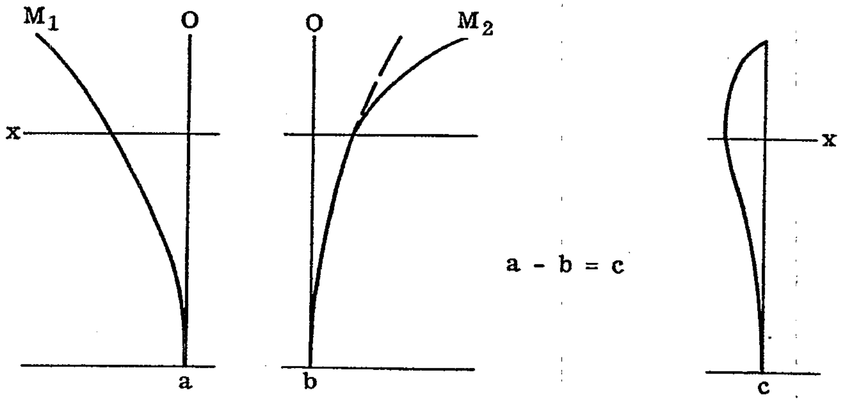


Figure 19.4- Plots of spherical aberration for the lens in Figure 19.3, with marginal aberration equal to zero.

the size of the air space.

19.2.5 Sphero-chromatism.

19.2.5.1 If the cemented doublet in Figure 19.3 is designed to eliminate spherical aberration for a wavelength in the middle of the spectrum, e.g., sodium D, we will discover that the C (red) rays will show considerable undercorrected spherical aberration, while the F (blue) rays will exhibit overcorrected spherical aberration. If the doublet is color-corrected to bring the paraxial C and F rays together, the spherical aberration curves will appear as in Figure 19.7 (a). Also, a better overall color correction will result if the paraxial rays are allowed to be slightly undercorrected, so that the two curves will intersect at approximately the 0.7 zone, or even a little higher as shown in Figure 19.7(b). The cause of this change in spherical aberration with wavelength will be made clear by the following discussion.

19.2.5.2 The refracting power of the doublet is very nearly the same for red (C), yellow (D), and blue (F). However, this refracting power is the sum of those of the positive and negative elements. If the lens is spherically corrected for yellow (D), the positive and negative elements will have certain refracting powers. For red (C) light, the individual refracting powers will be decreased (the sum remaining the same) because of the lower index of refraction. Conversely, for blue (F) light the elements' refracting powers will be increased. The nature of these changes are such, as to have the doublet spherically undercorrected for red (C) light and overcorrected for blue (F) light.

19.2.5.3 The air space can be used to correct this aberration. In Figure 19.8, white light entering a positive single element is refracted by the lens so that the F ray is bent more than the C ray. If the negative overcorrecting element is placed in contact with the positive element, no benefit is achieved. However, if the air space is present as illustrated in Figure 19.9, the F ray strikes the negative element at a lower height than the C ray and thus is subject to less overcorrection by this element. In this way, the naturally spherical undercorrection for red (C) and overcorrection for blue (F) can be neutralized or even reversed by increasing the air space. The amount of the correcting effect depends on the size of the air space and the angular interval between the (C) and the (F) ray upon emerging from the positive element. This angle depends upon the refracting power of the element and the dispersion of the glass.

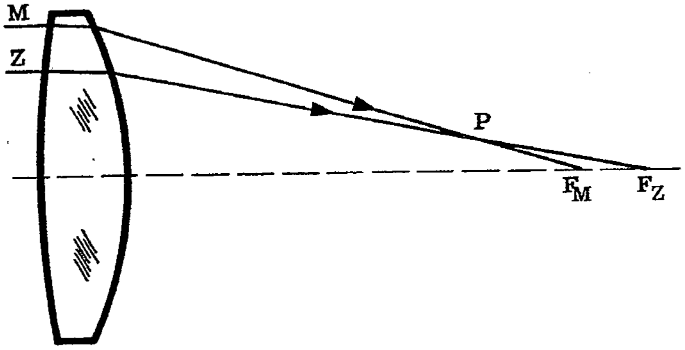


Figure 19.5- Single positive lens used in doublet of Figure 19.3.

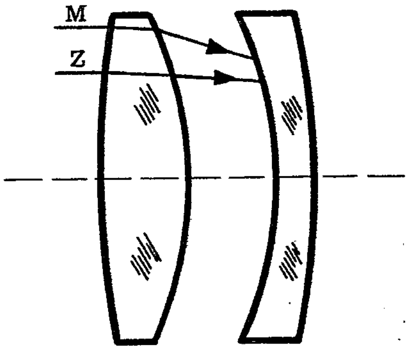


Figure 19.6 - Clark lens, illustrating air spacing to increase over-correction of zonal ray.

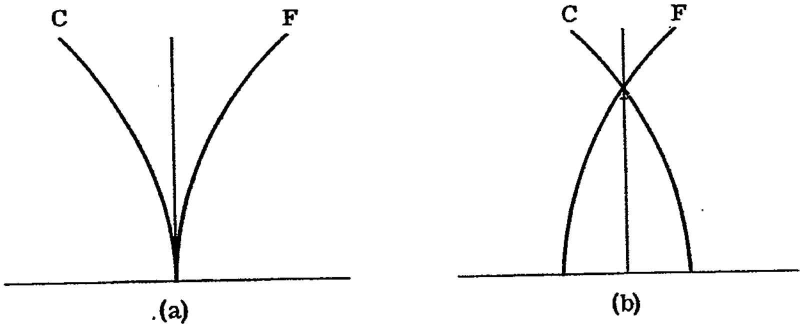


Figure 19.7- The doublet used in Figure 19.2 corrected for spherical aberration and color.

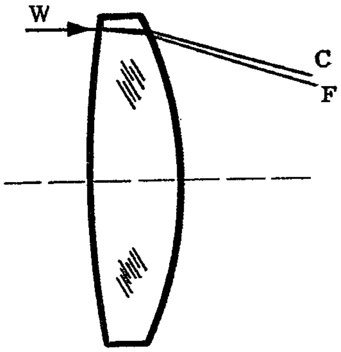


Figure 19.8- Refraction of monochromatic light by the positive element of the doublet shown in Figure 19.3.

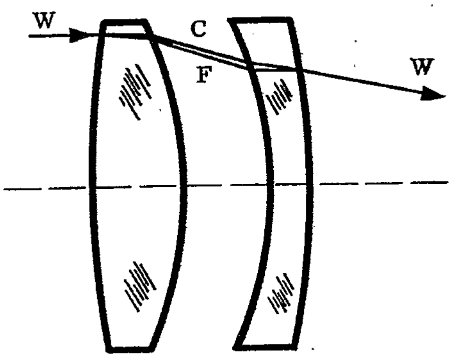


Figure 19.9- Refraction and correction of monochromatic light by the Clark lens shown in Figure 19.6.

19.2.5.4 The air spaced doublet shown in Figures 19.6 and 19.9 is generally known as the Clark lens, produced by Alvan Clark and his successors in the last half of the nineteenth century in the form of large aperture astronomical refractors of exceptional performance. However, there are several objections to the Clark lens. It is extremely difficult to correct both for zonal spherical and for sphero-chromatism with one and the same air space. Further, while in a cemented or contacted doublet the highly curved contact face is of low power, the separation into two elements means that the original contact face has become two strong opposing surfaces of high power. The result is that the axial adjustment and centration become quite critical and the lens become difficult to mount, adjust and maintain. These objections can be overcome by splitting the positive element into two parts and attaching one of these to the negative element as shown in Figure 19.10. Excellent performance with this combination can be achieved up to much higher apertures than is possible with either cemented or contracted doublets. See Section 11. The splitting up of the positive power into two elements means that all of the curves can be made quite mild, and additional degrees of freedom are available to the designer. The designer may choose the relative powers of the two positive components and, also, can investigate the possibility of making these components of different glass types, particularly, with respect to the dispersion factor.*

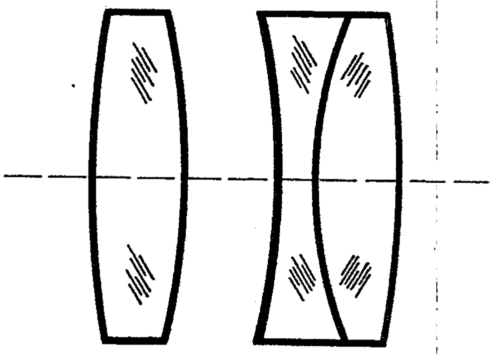


Figure 19.10-Modification of the Clark lens.

19.2.5.5 The lens form shown in Figure 19.11 has incorporated these principles into a simple and effective design used on one of the smaller theodolites. The actual construction has been used for a 40 inch f/6 lens and for a 24 inch f/5.6 lens. The front lens grouping incorporates the general principles discussed in the preceed-

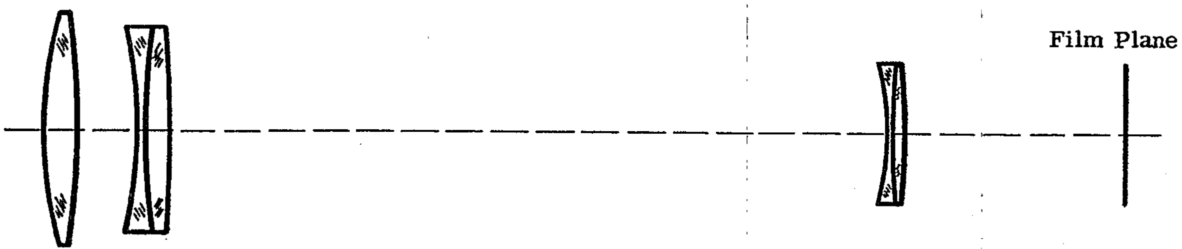


Figure 19.11- Optical layout, 24" high resolution lens.

---

* JOSA Vol. 42,7 pg. 451, S. Rosin.

ing paragraph. The rear negative doublet neutralizes the Petzval curvature and astigmatism of the front group so that high resolution is maintained over a 70 mm. film format. The spherical aberration curves for all visible colors are straight lines perpendicular to the axis, without any zonal bulge whatsoever (see Figure 19.12). This may be contrasted with the more usual type of curve such as shown in Figure 19.7b.

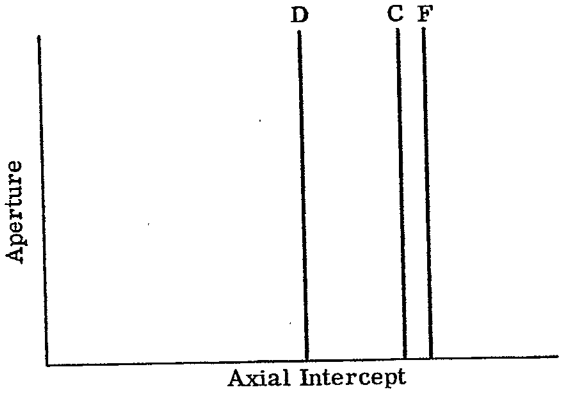


Figure 19.12- Spherical aberration curve for design shown in Figure 19.9.

19.3 REFLECTIVE SYSTEMS

19.3.1 Evolution of the reflective system.

19.3.1.1 The science of astronomy, based on the telescope as it is, discovered early in the twentieth century that the refractive objective had reached the limit of its development. To carry the physical apertures to ever higher values required the use of the reflective system, of which the principal converging element is the reflecting mirror. In this section, a few examples of advanced reflective designs will be given, but in order to properly orient the reader, and to enable him to effect his own designs for particular requirements, a short history of the evolution of reflecting systems and a discussion of some of their fundamental properties will be given.

19.3.1.2 The simplest of all of reflective systems is the concave mirror as shown in Figure 19.13. Rays from a distant object are converged by the mirror to a focus (F). If a film were placed at (F), an image at this point, which is located on the axis half-way between the instantaneous center of curvature (C) at the axis of the

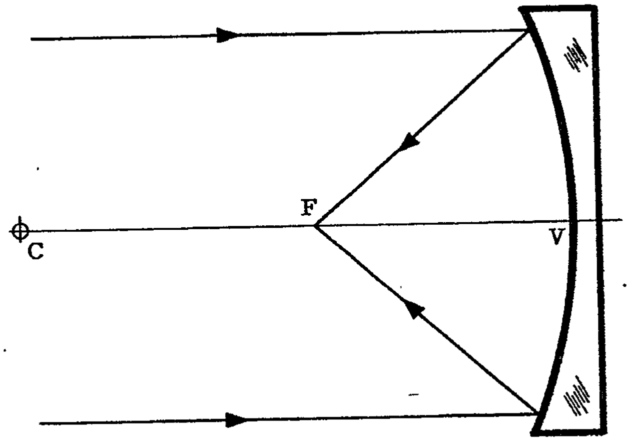


Figure 19.13- A concave mirror.

mirror and the vertex (V) of the mirror itself would be obtained. The simplest of all such mirrors is a portion of a sphere. However, a spherical mirror gives a very poor image at (F), since it is afflicted with a large amount of spherical aberration. This may be verified by ray trace, or by reference to the Seidel expression for spherical aberration. In order to sharpen the image at (F) in Figure 19.13, an aspheric surface for the reflector must be used, and this surface will be a paraboloid with its focus at (F).

19.3.1.3 The great astronomical instruments of the first half of the twentieth century fall into two main classes; the simple paraboloids, the greatest of which is the giant 200 inch diameter mirror at Mt. Palomar, and the Schmidt telescope.

19.3.2 The Newton system.

19.3.2.1 There is one serious mechanical disadvantage to the basic arrangement of Figure 19.13. If an eyepiece for visual observation, or a photographic plate is positioned at F, the center of the beam would be seriously interrupted. To overcome this difficulty, Newton, as early as the seventeenth century, proposed a plane mirror (M), to be positioned as shown in Figure 19.14, to bring the focus outside the beam where observation could be made. If a photographic motion picture film is placed at (F) in Figure 19.14, the recording apparatus can be made as large as necessary. A number of successful missile tracking devices have been constructed in accordance with this arrangement. It will be noted that the beam is partially obscured by the mirror (M). All reflective systems, except for the unimportant off-axis parabola, are characterized by this hole in the pupil, whether it is caused by a photographic plate or a mirror.

19.3.2.2 While it is true that a paraboloidal mirror forms a perfect image on the axis of the system, there remain important limitations. As has been discussed, physical optics and the finite wavelength of light impose a limitation on the resolving power of all optical systems. A good rule for this limitation is the simple equation

$$R = \frac{4.5}{a} , \qquad (2)$$

where R is the resolving power in seconds of arc, and a is the mirror aperture in inches.

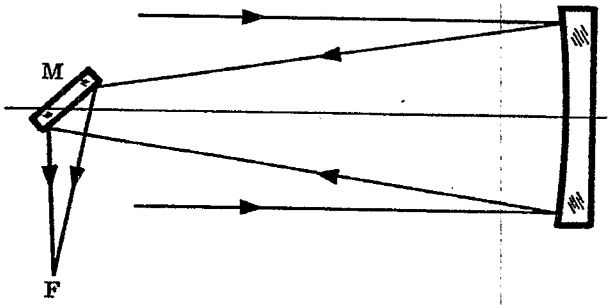


Figure 19.14- Newtonian mirror arrangement.

19.3.2.3 Another serious limitation on the properties of the paraboloid, is the presence of field aberrations, particularly coma. Indeed, the size of the useful field of view of a paraboloid, where the resolution over the field is in accordance with equation (2), can be given by

$$v = \left(\frac{f}{h}\right)^2 \Big/ 25 \qquad (3)$$

where v is the size of the field in inches, f is the focal length of the mirror, and h is the diameter of the mirror. The ratio $\frac{f}{h}$ is the f number of the mirror. Thus an f/5 mirror has a useful field of view of one inch, and an f/10 mirror has a useful field of four inches. The Mt. Palomar 200 inch diameter mirror works at f/3.3, so that its useful field of view is only 0.4 inch. However, over this 0.4 inch field, the Mt. Palomar paraboloid can theoretically resolve better than 0.025 seconds, which for a focal length of 660 inches, would amount to 0.00008 inch.

19.3.3 The Cassegrain system.

19.3.3.1 Another class of reflecting system is the two mirror Cassegrain arrangement.* This system is extremely popular in missile tracking instruments and is shown in Figure 19.15. Rays from a distant object strike a concave mirror (M_1) and are reflected towards a focus at (F_1). Before the rays are converged, a second mirror (M_2), which is convex, interrupts the beam and reflects it to a second focus at (F_2). The position of (F_2) outside the system puts it in an extremely convenient and favorable position for image recording. The hole in the mirror (M_1), is in the region which is blocked out of the original bundle by the physical presence of the convex mirror (M_2). The convex mirror is supported by a mechanical spider, or is cemented to a flat glass plate. In all cases, the mirror (M_2) magnifies the image considerably, since the distance from (F_1) to (M_2) is considerably less than from (M_2) to (F_2). A favorite value for this ratio is 4x, although this figure may vary considerably. This factor lengthens the focal length over that of the mirror (M_1) by the same amount, and similarly increases the focal ratio or f/#. Thus, the Cassegrain arrangement is very suitable for systems of long focal length and relatively low illumination.

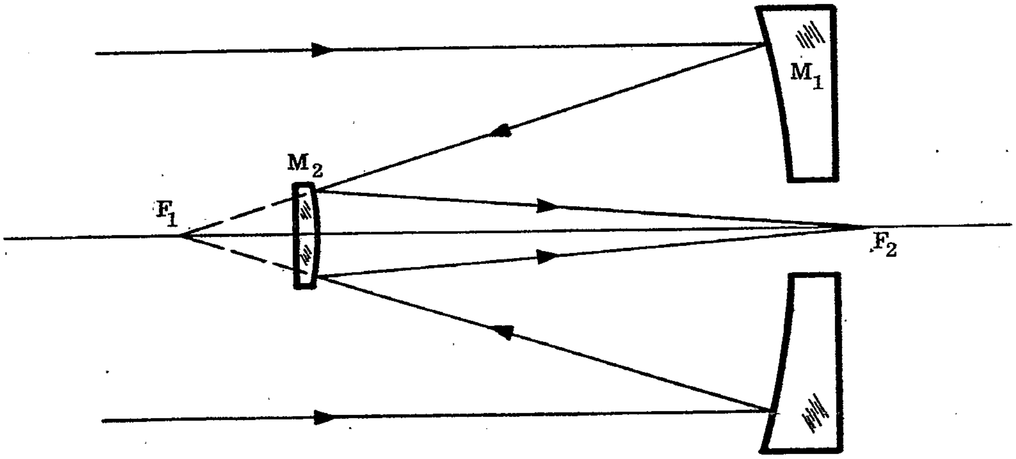


Figure 19.15- Cassegrainian Mirror arrangement.

19.3.3.2 However, there is always the requirement that the image at (F_2) be sharp, that is to say, free from spherical aberration. There are an infinite number of combinations of the two surfaces which will achieve this. For example, the mirror (M_1) can be spherical, in which case (M_2) is a complex, higher order curve. Conversely, the mirror (M_2) can be spherical, in which case the mirror (M_1) will be of a complex nature. Or, in its favorite form, (M_1) is a paraboloid with focus at (F_1), and (M_2) is a hyperboloid with foci at (F_1) and (F_2). Which of these combinations is the best? In most cases, the arrangement which gives freedom from coma will be the most desirable.

---

* While the system originally proposed by Cassegrain consisted of a paraboloidal primary with an hyperboloidal secondary, accepted usage today has broadened the term "Cassegrain" to apply to any system consisting of a concave primary and a convex secondary.

19.3.3.3 A power equation may be set up for (M_1) with

$$-x = a_1 y^2 + b_1 y^4 + c_1 y^6 + \dots\dots , \tag{4}$$

and the equation cut off after the first two terms. For all surfaces

$$a_1 = 1/2\, r_1 , \tag{5}$$

where r_1 is the instantaneous radius at the vertex of (M_1). Then let b_1 assume a continuous set of values in the equation

$$-x = y^2/2r_1 + b_1 y^4 \tag{6}$$

and for $b_1 = 0$, the surface is a paraboloid. For $b_1 = 1/8\, r_1{}^3$, the surface is spherical, provided the semi-diameter of the mirror is not too large a fraction of the radius r_1.

19.3.3.4 A similar equation can be set up for the mirror (M_2) with

$$-x = y^2/2r_2 + b_2 y^4 \tag{7}$$

where r_2 is determined by the positions of the paraxial arrangement (F_1), (F_2), and (M_2), and b_2 is determined by requiring the marginal ray to pass through (F_2). The form of the mirror (M_2) for any given value of b_1 can then be found. The calculation is difficult and can best be effected on an electronic calculator if available. Then one paraxial and one marginal ray can be traced for each combination (the paraxial ray can be the same for all the combinations) and the departure from the sine condition can be determined. In this procedure, it will be found that the combination most nearly free from coma is not too far from the paraboloid-hyperboloid arrangement, in cases where the magnification of (M_2) is not too different from 4x. If complete freedom from coma is desired by the designer, some departure from this combination may be indicated. However, in all designs known to the author, the paraboloid - hyperboloid form is used, except for extremely low aperture systems, where the spherical aberration is unimportant, and two spheres may be employed.

19.3.3.5 Up to this point, the simple mirror arrangements have been discussed chiefly in the form of the paraboloid and the Cassagrain two mirror system. No color aberrations are involved in purely reflective systems. After spherical aberration is corrected, the paraboloid affords no more degrees of freedom to correct coma. In the Cassagrainian arrangement, the proper choice of form allows both spherical aberration and coma to be controlled. However, no mention has been made of the remaining aberrations, namely astigmatism, field curvature and distortion. These aberrations are handled in precisely the same way as in refractive systems. There is one important difference, relating to the Petzval field curvature.

19.3.3.6 It will be recalled that in refractive systems with a large excess of positive power, the Petzval curvature is concave towards the incident light. The reverse is true in reflective optics. A converging element (concave mirror) has associated with it a heavy Petzval curvature convex towards the incident light. This affords the possibility of combining converging reflective and refractive systems to achieve a flat field, as shall be discussed later. If r is the radius of the mirror, and c ($= 1/r$) its curvature, the contribution to Petzval curvature of any reflecting surface is given by

$$P = 2Nc \tag{8}$$

where N is the index of refraction of the medium in contact with the mirror. For a single concave mirror, the Petzval surface is concentric with the mirror surface as shown in Figure 19.16.

19.4 CATADIOPTRIC SYSTEMS

19.4.1 Introduction. Now consider the second class of reflective systems, which include the Schmidt arrangement and some of its variations and developments. There were two fundamental principles discussed in previous sections of this handbook relating to the change in aberration with a shift in the stop position. In brief, these principles are,

(1) The change in the Seidel coma coefficient, due to a change in the position of the stop, is proportional to the spherical coefficient, multiplied by the shift in the stop position.

(2) The change in the Seidel astigmatic coefficient is proportional to the coma coefficient, multiplied by the shift in the stop position.

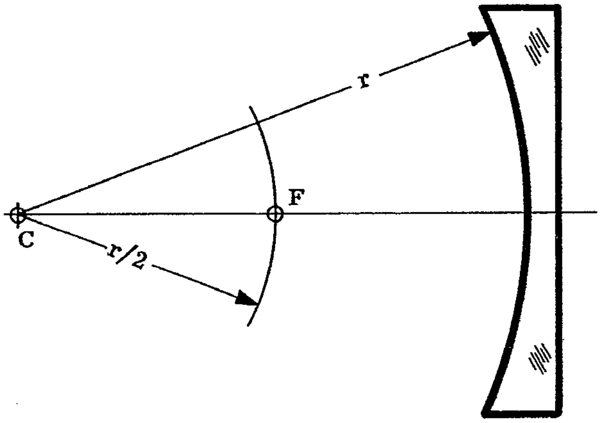


Figure 19.16- Curved focal surface, concave mirror.

A very important principle, relating to aspheric surfaces may then be added,

(3) Any change in form, without a change in the vertex radius, has no effect on the off-axis Seidel coefficients for an optical surface located at the stop position.

The third principle has far reaching consequences in the design of optical systems, since it allows the optical designer to correct spherical aberration by aspherizing the surface at the stop position. At this point, it will only have a mild effect on axial color (no effect at all in the case of reflecting systems), and no effect on coma, astigmatism, field curvature, distortion, and lateral color. It is not to be implied that the optical designer need limit the aspherizing of elements to the stop position, since the effects of such a procedure are subject to calculation, but the use of such surfaces in positions other than the stop brings them into the general juggling procedure characteristic of optical design.

19.4.2 The Schmidt system.

19.4.2.1 With respect to the contributions of Schmidt to optical science, it will be recalled that the single paraboloidal reflector had no spherical aberration but an extremely large amount of coma. In accordance with principle 1, the stop can be anywhere, without changing this coma, since the spherical aberration of the paraboloid is zero. If the stop is considered to be at the mirror, which would be the case if there are no artificial diaphragms in front of it, and the form of the mirror is allowed to change to something else, principle 3 states that the off-axis aberrations, including coma, will be unaffected. Indeed all concave mirrors of equal vertex radius, be they paraboloid, sphere, hyperboloid or off-beat curve will have equal amounts of coma, astigmatism and field curvature. The Seidel coefficient for astigmatism is equal to the refracting power as it is in the case of all thin systems at the stop. For the simple mirror, this is equal to 2c.

19.4.2.2 Now consider the apparently strange consequence of principle 3, and also consider principles 1 and 2. The optical designer might say that perhaps the large amounts of off-axis aberrations can be reduced by shifting the stop position. The designer will be frustrated in the case of the paraboloid, since there is no change in coma due to the absence of spherical aberration. He will probably choose the sphere, since as an accomplished technician he knows there will be less difficulty in making the sphere, than is the case with the hyperboloid or off-beat curve. If the designer allows the stop to recede from the sphere, he will notice a decrease in both coma and astigmatism, until the center of curvature of the mirror is reached. At the center of curvature, these aberrations will vanish and the configuration will be as shown in Figure 19.17.

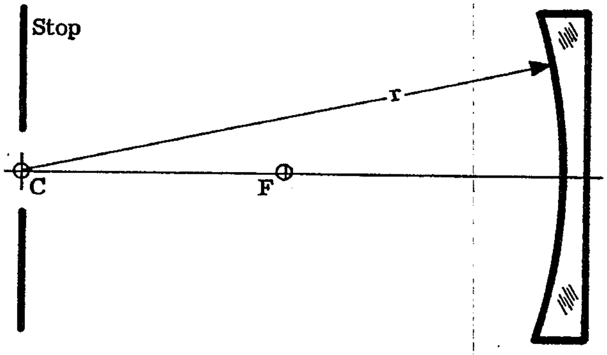


Figure 19.17- Schmidt arrangement, correction plate not shown.

19.4.2.3 At this stage, the design is free of coma and astigmatism, but still contains a large amount of field curvature, which must be tolerated, and a considerable amount of spherical aberration. Principle 3 will allow for the correction of the latter, by placing a plane parallel refracting plate at the stop at c in Figure 19.17, and deforming one surface to correct the spherical aberration. Since the spherical mirror undercorrects the marginal ray, the edge of the plate must have a slight negative power to neutralize it, as compared to the center. In practice, it is common to impart a tiny central positive power to the plate, resulting in a parallel zonal section as shown in Figure 19.18, and a reduced negative marginal region, thus shortening the overall focal length by a very small amount. With the parallel zonal section approximately 0.8 of the distance from the center to the margin, the plate imparts the minimum axial color to the beam.

19.4.2.4 It is fairly certain that Schmidt did not follow this rather involved reasoning to conceive his system. It is quite obvious that for a stop placed at the center of curvature of a spherical mirror, the coma and astigmatism must be zero, since the chief rays strike the mirror normally and can define a new axis just as valid as the central axis. There can be no coma or astigmatic difference on the axis of an optical system. However, the more involved reasoning first given is capable of further extension and application as shown below in the case of the oblate spheroid.

19.4.2.5 The Schmidt telescope has the very great advantage over the paraboloid of enormously extending the field of view over which the image remains sharp. The largest made to date is located on Mt. Palomar, and has a 48 inch diameter aperture with an ability to take sharp pictures over an area 14 inches square. The plates must assume the shape of the field curvature. The focal ratio is F/2.5 with a focal length of 120.9 inches. However, the Schmidt suffers from other defects. Since it has its corrector at the center of the curvature of the mirror, the system must be twice as long as its focal length. Also, to prevent vignetting, the primary mirror must be considerably larger than the aperture. In the Mt. Palomar instrument, the spherical mirror is 72 inches in diameter.

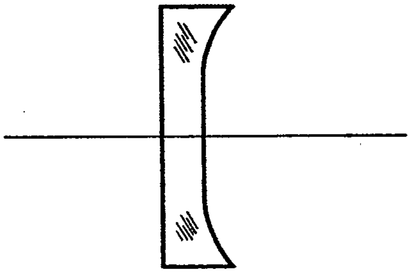
19.4.2.6 If only moderate extension of the field of view is desired, the Schmidt type arrangement, with an oblate spherical primary, can achieve this in a much shorter structure than the Schmidt with a spherical primary. In the following equations,

$$-x = \frac{1}{2r} y^2 , \qquad (9)$$

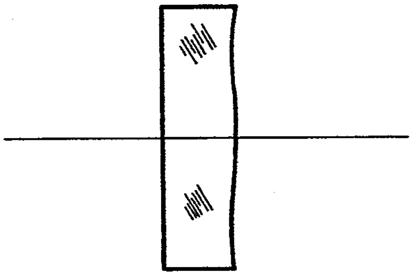
$$-x = \frac{1}{2r} y^2 + \frac{1}{8r^3} y^4 , \tag{10}$$

$$-x = \frac{1}{2r} y^2 + \frac{2}{8r^3} y^4 . \tag{11}$$

Equation (9) is the equation of the paraboloid, equation (10) represents a sphere (expansion of its equation to two terms), and equation (11) represents a surface twice as heavily curved away from the paraboloid as the sphere. Any surface in which the coefficient of the y^4 term is in excess of $1/8r^3$ is called an oblate spheroid and equation (11) defines one such surface.



Schmidt with parallel central section.



Schmidt with parallel zonal section.

Figure 19.18- Schmidt plate with parallel central section, (a); and parallel zonal section, (b).

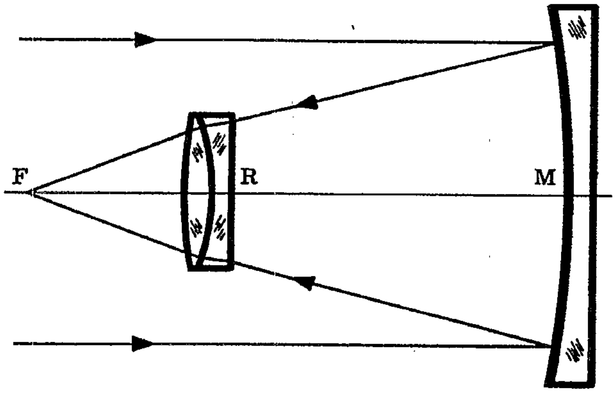
19.4.2.7 The Seidel theory of aberrations as applied to aspheric surfaces states that if equation (9) defines, as it does, a surface with no spherical aberration, and equation (10) defines a surface with a certain amount, then equation (11) is a surface with twice the spherical aberration of that defined in equation (10). Referring back to principle 1, it will be recalled that spherical aberration is needed in order to correct coma by shifting the stop, and that it is possible to correct the coma of the sphere by shifting the stop from the mirror to the center. If now, the designer starts with a surface in accordance with equation (11) having twice the spherical aberration of the sphere, we need to shift the stop only back to the focus to correct the coma, shortening the instrument to half of the Schmidt. The correction plate will now have to be made to correct the doubled spherical aberration.

19.4.2.8 The situation with respect to astigmatism is not so fortunate. Since the stop-at-mirror coma is identical for equation (10) and (11), principle 2 states that shifting the stop back to the focus will correct only half of the astigmatism. For maximum field the stop may be shifted a little further, thus reducing the astigmatism, but allowing the coma of opposite sign to creep back in, until a desirable compromise is reached. The Schmidt principle is capable of a number of variations which will not be discussed further.

19.4.3 The Ross-Baker system.

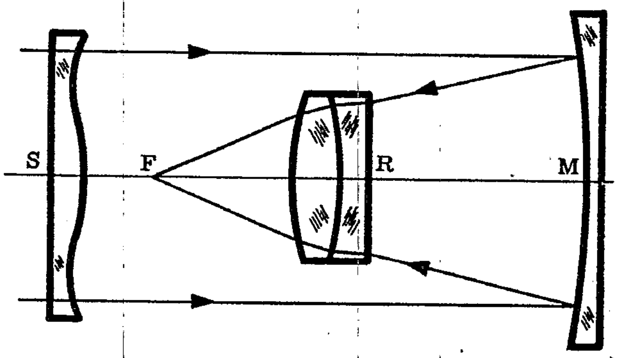
19.4.3.1 Efforts to extend the field of view of the paraboloid have met with some degree of success. Ross constructed some lenses spaced close to the focal plane in accordance with the arrangement in Figure 19.19. A color corrected doublet, placed at R as shown in Figure 19.19, is given sufficient power so that its positive Petzval field curvature contribution will just neutralize that of the mirror M. Its bending and spacing from F offer two degrees of freedom, which are used for the correction of coma and astigmatism. Unfortunately, it proved impossible to prevent R from reintroducing undercorrected spherical aberration into the system so that the images tended to be soft. However the field of view of the paraboloid was greatly extended by this maneuver.

19.4.3.2 James G. Baker has proposed the addition of a Schmidt corrector plate to the Ross system as shown in Figure 19.20. The combination of M and R in Figure 19.20 is designed for correction of coma and astigmatism for a stop position at S. Then insertion of the Schmidt plate takes out the residual spherical aberration. An analysis of this system shows that before the insertion of the Schmidt plate, the residual spherical aberration amounted to a substantial fraction of that of a spherical mirror. However, the amount of depth needed to be hollowed out of the Schmidt was only a few wavelengths of light. This illustrates the tremendous leverage exerted on the rays by systems of this type, and the demanding exactitude required for their construction. Baker proposed this system, which gives definition over a field of view an order of magnitude larger than that of the paraboloid, as a means of correcting simple paraboloids already in existence by the addition of R and S. A similar requirement makes this system desirable for missile recording. The complete system can be used for photographic tracking, while the removal of the refracting elements S and R allows conversion of the system for the detection of targets in the medium infra-red and in the ultraviolet, where the glass would be opaque.



NOT TO SCALE
The size ratio of element R to element M
is approximately 5 to 1

Figure 19.19- Ross arrangement.



NOT TO SCALE
The size ratio of element R to element M
is approximately 5 to 1

Figure 19.20- Baker arrangement.

19.4.4 Modification of the Ross-Baker arrangement.

19.4.4.1 The author of this section has proposed, in an exchange of letters with Dr. Baker, the elimination of the correcting plate S and transferring its function to the mirror M, thus deforming the paraboloid. The combination of deformed paraboloid, and Ross lens R would have the disadvantage of making the mirror unusable by itself. However, it would have the advantage of eliminating the only large refracting element and enabling the system to be carried to higher physical dimensions.

19.5 APPLIED SYSTEMS

19.5.1 Satellite tracking camera.

19.5.1.1 The principles on which the design of the more complex optical systems used in missile and satellite tracking are based, are illustrated in the Figure 19.21. The design is a classical Schmidt system with just a few variations. For the purpose intended, this camera was designed for high light gathering power and large field, particularly in one direction, preferably the direction of satellite path. The physical aperture of the system is 20 inches, and with a focal length of approximately the same value, the system operates at f/1. To prevent vignetting the primary spherical mirror is 31 inches in diameter.

19.5.1.2 It will be noted that the aperture of the system is very close to the center of curvature of the primary mirror, but the single correcting plate, which normally is located there, is split up into a color corrected triplet for the purpose of eliminating the small amount of residual axial color in the single Schmidt plate. The four inner surfaces of this system are aspheric.

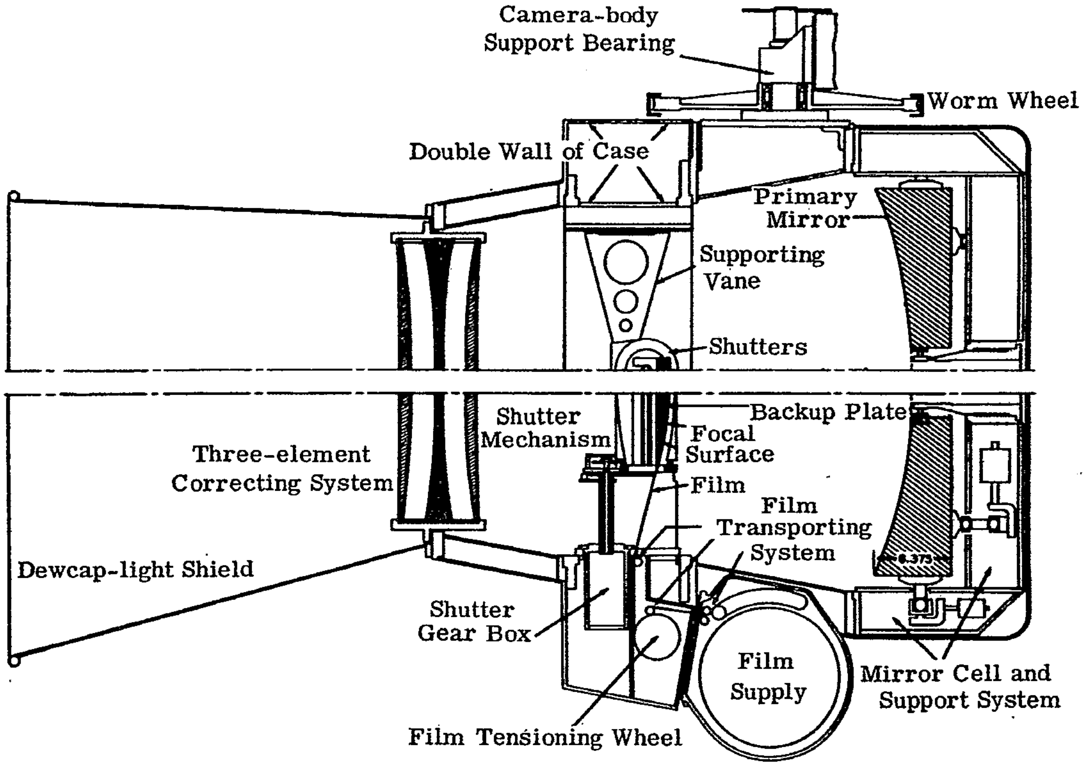


Figure 19.21-Section view of a satellite-tracking camera. (Courtesy of Perkin Elmer Corp.)

19.5.1.3 It is presumed that, because of the high relative aperture of this system (f/1) the curvature of the Schmidt plate required would be more extreme than usual, leading to more axial color than the designer could tolerate. The splitting up of the single plate into three, with the central glass different from the outside, and the distribution of the Schmidt curvature among four surfaces would tend to alleviate this situation.

19.5.1.4 It will be noted from Figure 19.21, that the film is transported over a spherically curved gate, which matches the curved focal plane of the image. The curvature in the plane at right angles must necessarily be zero, because of the mechanical impossibility of bending the moving film into a compound curve. Consequently the field coverage in this direction is limited to only 5 degrees, while in the direction of film travel it reaches the amazing value of 31 degrees. It was found, that at the edges of this extreme field the focal surface departs slightly from a spherical shape so that the film runners are not quite circular. The combination of careful design and excellent execution resulted in a system wherein 80 percent of the point energy anywhere in the field is within a circle 0.001 inch in diameter. This instrument was conceived for the purpose of tracking the U. S. Vanguard satellite, and the first instrument arrived just in time to be used for the original Sputniks.

19.5.2 ROTI Mark II (Recording optical tracking instrument).

19.5.2.1 This instrument and the Igor were made to similar specifications but each has features worth discussing. The original requirements envisioned a versatile instrument capable of a series of fixed focal lengths ranging up to 500 inches in value. Another requirement was that the instrument could be adapted for infrared which necessitates the choice of a paraboloid for the primary mirror.

19.5.2.2 It would be thought that some form of the reflector-corrector system of Baker (Figure 19.20) would be indicated and such is indeed the case. However, the corrector feature is introduced in a rather unique fashion.

19.5.2.3 Referring to Figure 19.22, light enters from the left, passes through the window and strikes the primary mirror, a paraboloid of 100 inch focal length and 24 inch aperture. After reflection by two Newtonian mirrors as shown, the rays are brought to a focus at the reticle. Just before this point a pair of sliding wedges introduce a variable amount of glass into the path. By adjustment of these wedges, the instrument focus can vary between 3000 yards and infinity, the focus being automatically controlled by range data determined from the associated radar equipment.

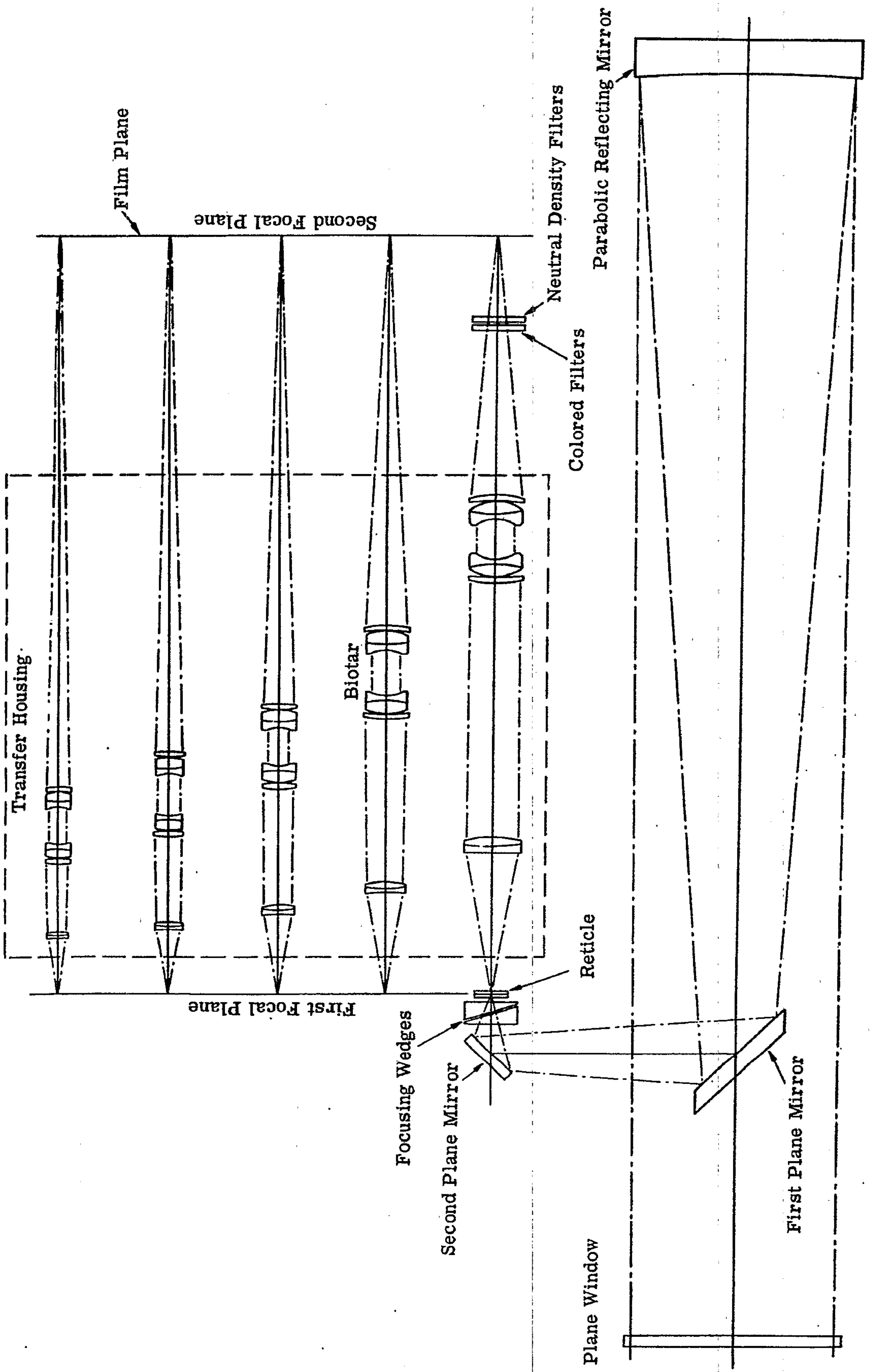


Figure 19.22– R. O. T. I. Mark II, Optical arrangement. (Courtesy of Perkin Elmer Corp.)

19.5.2.4 As is known from the previous discussion, the primary image is heavily afflicted with coma, and the astigmatism is only partially corrected by placement of the stop at the window. The doublet lens, placed behind the focal plane, serves three functions. First it acts as a Ross lens to neutralize the coma resulting from the paraboloid; second, it is a collector lens to turn the rays into the reimaging Biotar system; finally it imparts an initial magnification (2x) to the image at the prime focus.

19.5.2.5 The five imaging systems can be inserted into the system at will. These have magnifications ranging from 1x to 5x in unit power steps. With the doublet lenses working at 2:1, the Biotar type imaging lenses work at magnifications of 1/2, 1 . . . . 2-1/2x.

19.5.2.6 Where is the Schmidt type correcting surface characteristic of the reflector corrector design? Normally, it would be at approximately the location of the window. However a conjugate stop position exists in this system, namely the midpoint of the Biotar imaging lens. The aspheric correction required is put on the two innermost surfaces of this lens, those adjacent to the aforementioned stop position. This aspheric correction is for the purpose of eliminating the zonal spherical aberration only, since the primary spherical is taken care of in the design of the refractive elements. KZF Schott glass is employed in the system to reduce considerably the secondary spectrum.

19.5.2.7 After passing through the imaging lens, the rays traverse filters, colored or neutral as desired. The latter are automatically controlled by photoelectric means. The final focal plane is at the gate of the 70 mm camera, with a 2-1/4 inch square field of view.

19.5.2.8 Recapitulating, the five systems give a range of focal lengths from 100 inches F/4 to 500 inches F/20 (approximate figures, not allowing for the occlusion by the first Newtonian mirror). When enough light is available, the 500 inch focal length reaches into extremely long distances for the target missile, and the writer believes this has proved the most used focal length. Tracking is effected by means of two operators, one for azimuth and one for elevation, with appropriate telescopes. An elaborate electrical control system is very effective in the accuracy of tracking. A photograph of ROTI is shown in Figure 19.1 .

19.5.3 Igor.

19.5.3.1 This instrument is somewhat smaller than ROTI but serves essentially the same purpose, i.e., the observation of missiles in flight to extreme distances. The original specifications envisaged some situations where a 70 mm camera (2-1/4" x 2-1/4") was to take pictures under some circumstances at the prime focus (P) as shown in Figure 19.23. Consequently the image at this point had to have complete correction for all aberrations over its 2-1/4 inch by 2-1/4 inch format. The combination of Schmidt Plate (S), primary mirror (M), Newtonian mirrors (N1) and (N2), and Ross lens (R) as shown in Figure 19.23 accomplished this purpose very well. Removal of (S) and (R) allows for infrared and ultraviolet measurements if desired. Focussing of the system from infinity to 3000 yards is effected by motion of the Ross lens (R) towards the mirror (N1). The total travel to cover this range is 3/16 inch.

19.5.3.2 The system works at F/5 with a clear aperture of 18 inches, so that the focal length of the system at prime focus is 90 inches. The focal length of the paraboloid itself is 118 inches, reduced to the required 90 inches by the Ross lens. A collector lens at (P) and a re-imaging system at (L) enables final imagery on a 70 mm camera at (F).

19.5.3.3 For a change to longer focal lengths and magnified images a system of Barlow lenses is employed. A Barlow lens is illustrated in Figure 19.24. Suppose rays are converging to a prime focus at P. A negative lens B is inserted into the path and diverges the rays to a more distant focus at F. The image is magnified by the ratio BF/BP.

19.5.3.4 Referring to Figure 19.23, the 1x system of collector and re-imaging lenses are removed and one of the Barlows, B_2 (2x), B_4 (4x) or B_5 (5.7x) is inserted, transferring the final image to the camera at (F) under the particular desired magnification. The highest available magnification corresponds to a focal length in excess of 500 inches.

19.5.3.5 The variable density filter arrangement at (D) is comprised of two oppositely rotating continuously varying density filters, slightly inclined to each other and to the axis, to eliminate multiple reflections. Two are needed to keep the field uniform. As in ROTI, the image is quite good over the field of view in all powers.

19.5.4 S.M.T. (Small Missile Telecamera).

19.5.4.1 The requirements for this instrument are similar to those for the satellite tracking camera except that a relatively small field of view is required, namely that of a 70 mm camera. High light gathering power (low F/#) is indicated for the tracking of small, high velocity missiles. The focal length of the system is 100 inches with a 30 inch aperture, so that the system is working at F/3.3, and the field is 1° 18'. This system

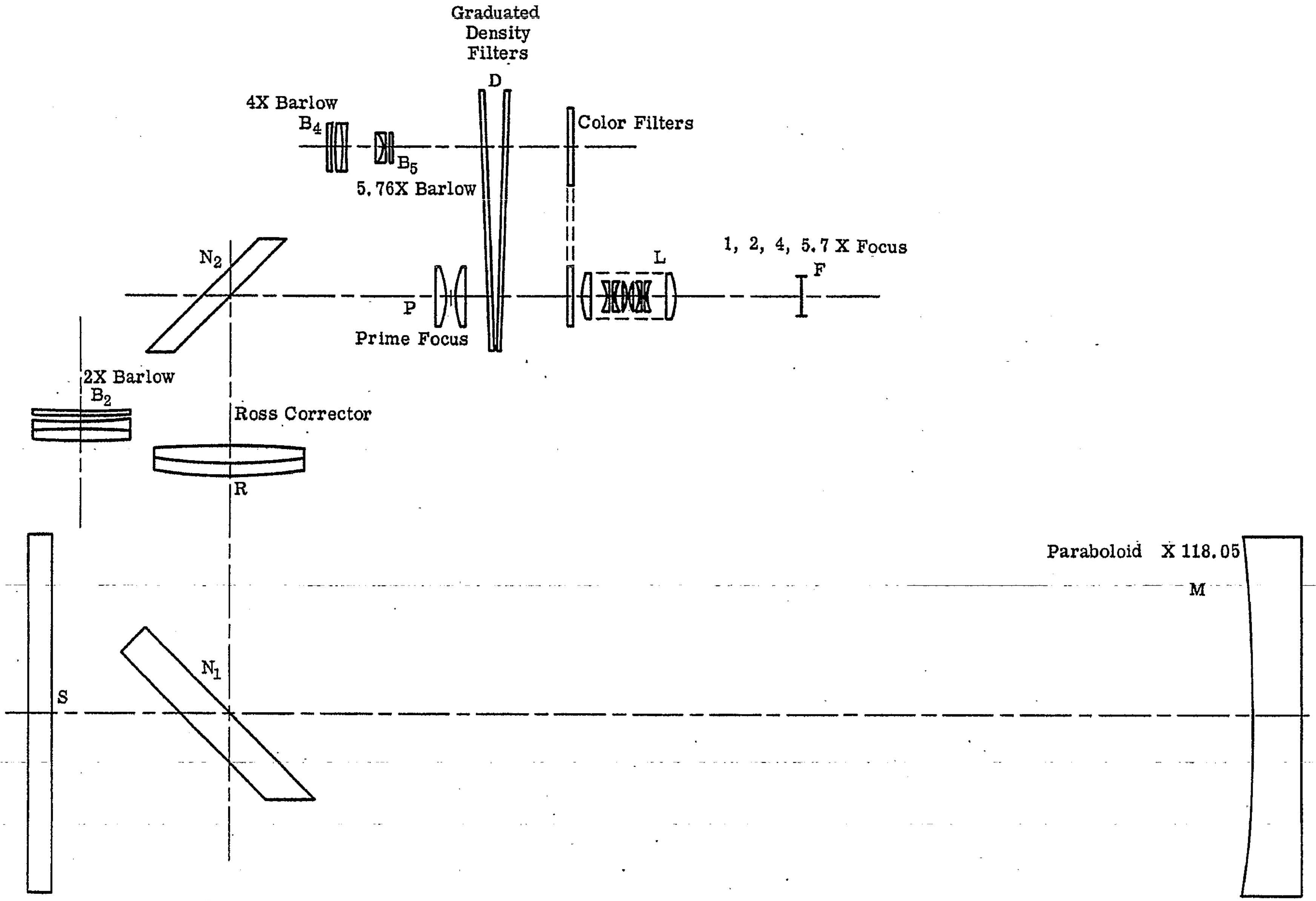


Figure 19.23-I.G.O.R. Optical arrangement. (Courtesy of American Optical Co. Drawing No. 013-0032)

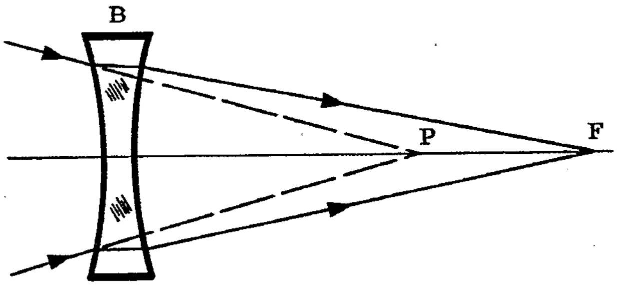


Figure 19.24- A Barlow lens.

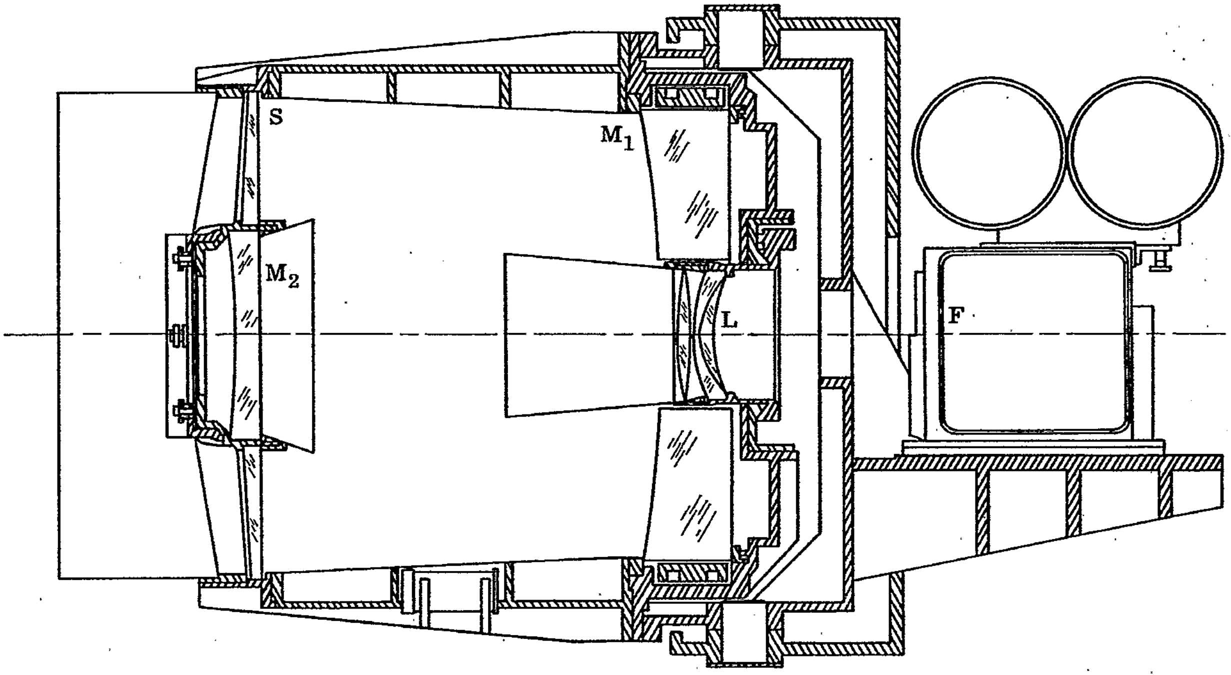


Figure 19.25- S. M. T. Optical arrangement.
(Courtesy of Perkin Elmer Corp.)

is similar to the Cassegrain arrangement in that two mirrors are involved, with the final image behind the hole in the primary as shown in Figure 19.25.

19.5.4.2 Light enters from the left and strikes the spherical mirror (M), after passing through the Schmidt plate (S). The convergent beam is interrupted by the mirror (M_2), which is actually a lens with its rear surface reflecting (often known as a Mangin mirror). The beam thus reflected a second time traverses the lenses placed in the central hole in the primary before reaching the film plane at (F).

19.5.4.3 This arrangement delivers an excellent flat field over its 2-1/4 inch x 2-1/4 inch field. The Mangin mirror and hole lenses afford enough degrees of freedom to correct for the coma and astigmatism of the shortened Schmidt arrangement, as well as to balance out the color aberrations they themselves bring in. The missile trajectory is predicted beforehand and the instrument is aimed from stellar observations made with a calibration camera.

19.5.4.4 It will be noted that heavy emphasis has been given to the optical arrangement of these complex instruments and their illustration of the optical principles discussed earlier. No space is available here for the description of their complex mechanical, electrical and electronic construction which tax all the resources of modern technology.

20 APPLICATIONS OF THIN FILM COATINGS

20.1 INTRODUCTION

20.1.1 General uses of thin films. Quite frequently optical components are coated with thin layers of various solid materials for the purpose of altering either their physical or optical properties. As an example of the former purpose, aluminum mirrors are often coated with a thin layer of silicon monoxide in order to increase their resistance to abrasion and chemical attack. The addition of this layer alters the spectral reflectivity of the mirror, although this is not the primary purpose of such a coating. More frequently, however, thin film coatings are used for the primary purpose of altering the spectral reflection and transmission of optical components. Sometimes a thin film coating consists of only one layer deposited upon a suitable substrate. In other cases, many layers, often as many as forty or fifty, are used to produce a given optical filter. Hence this type of filter is called a multilayer filter, or simply a multilayer. In Section 20, the term multilayer is used as a generic term for such thin film coatings, even though the "multilayer" coating may consist of a single-layer antireflection coating on a glass surface.

20.1.2 Typical applications.

20.1.2.1 Antireflection coatings. Whenever light traverses an interface between two media of different refractive index, such as an air-glass interface of a lens, some of the light is reflected. Often the spacing of optical elements is such that these reflections are manifested in the image plane as "flare images". Before the advent of antireflection coatings, many otherwise acceptable lens configurations were rejected because they produced these flare images. The coating of optical surfaces with antireflection coatings has practically eliminated this problem. It is important that antireflection coatings be applied to infrared optical components, such as lenses or domes, which contain germanium, silicon, or other materials with a high refractive index. The loss of light at uncoated surfaces would be prohibitive otherwise. Antireflection coatings are discussed in more detail in Section 20.3.

20.1.2.2 Achromatic beam splitters. Many optical devices, such as interferometers, range finders, optical gunsights, utilize beam splitters which divide a light beam and divert it into two directions. Thin metal films have been used as beam splitters for many years, but they are inefficient because the metal absorbs part of the light. More recently, multilayer beam splitters have been developed which are much more efficient, because they contain only non-absorbing materials. Less than one percent of the light is absorbed in a typical multilayer beam splitter; the remaining 99% of the light is either reflected or transmitted. The properties of multilayer beam splitters are expounded in Section 20.7.

20.1.2.3 Color filters and band-pass filters. Multilayer filters are used to transmit (or "pass") a broad band of wavelengths in one spectral region, but attenuate in other regions. For example, a multilayer filter is available which transmits more than 90% in the blue but has a transmission of less than 0.5% in the green and red. This multilayer filter is superior to the conventional glass or dyed-gelatin absorption filters, which have a much lower transmission in the blue. Similar types of band-pass filters have been developed for the ultraviolet and infrared spectral regions. The spectral transmission of some typical multilayer filters is shown in Figures 20.86, 20.91 and 20.92; a general discussion is given in Sections 20.5.2 and 20.6.2.

20.1.2.4 Color-selective beam splitters. Figure 20.1 shows a multilayer which is used as a color-selective beam splitter. In this example, the beam splitter transmits blue light, but reflects the green and red. Such beam splitters are useful as color separation devices in color photography and color television. This type of beam splitter is often called a dichroic mirror; its properties are explained in Section 20.7.2.

20.1.2.5 Narrow pass-band (interference) filters. Multilayer filters are used to transmit a narrow band of wavelengths. For reasons which are described in 20.10.2.2, these narrow-band filters are called interference filters, although in a technical sense all multilayer filters are interference filters, because they depend upon the interference of light reflected from the various films. One type of interference filter which is manufactured commercially in large quantities has a pass band which is from ten to twenty millimicrons wide in the visible spectral region. Custom-made filters have been produced which have a pass band as narrow as 0.1 mμ. A filter of this type has been used to isolate one of the sodium D lines at 589.0 mμ from its neighbor at 589.6 mμ. Such filters have many potential uses in the field of spectro-chemical analysis and are discussed in more detail in Section 20.10.

20.1.2.6 Semi-transparent mirrors. Multilayer mirrors have been produced which not only have a high reflectivity, but also transmit almost all of the light which is not reflected with a small absorption loss. A typical multilayer mirror might reflect 95% of the incident light and transmit 4.5%, the remaining 0.5% being absorbed or scattered. These multilayer mirrors have a much lower (absorption) loss than the conventional semi-transparent films of silver or aluminum and are useful for coating the plates of a Fabry-Perot interferometer or the ends of an optical maser. The spectral reflectivity of a semi-transparent metal mirror

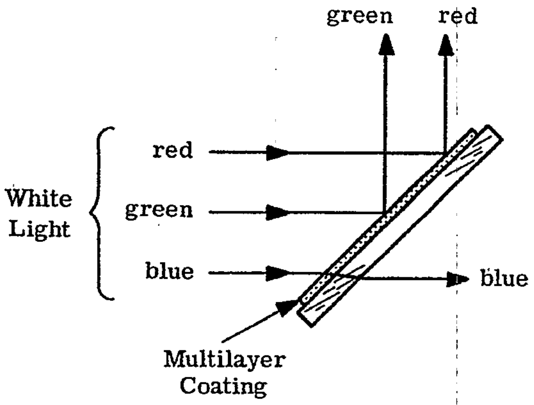


Figure 20.1-A color selective beam splitter which reflects the red and green, but transmits the blue.

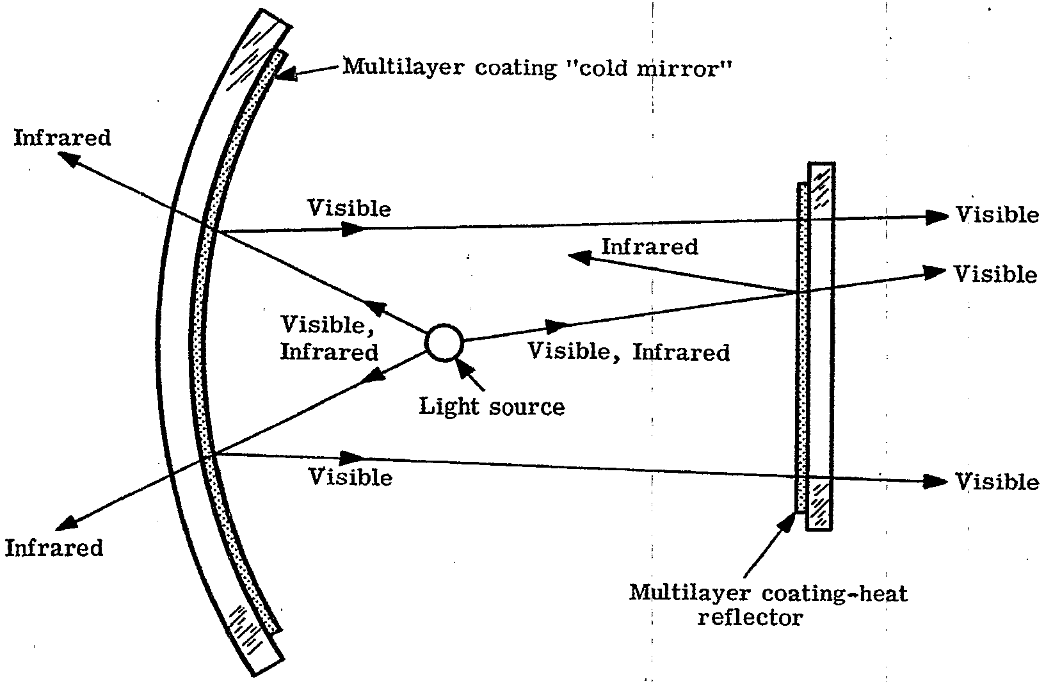


Figure 20.2-A cold mirror and heat reflector reduce the heat directed towards the film in a projection system.

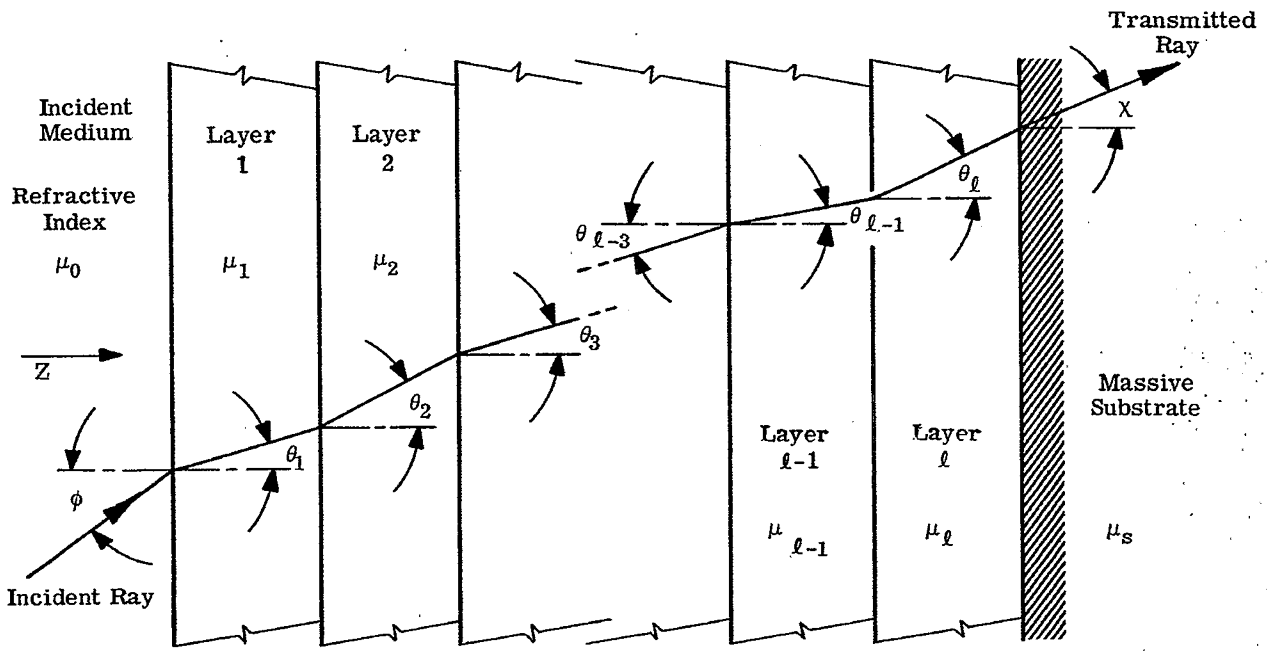
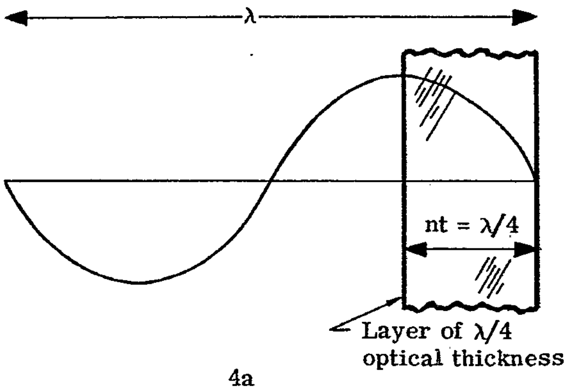
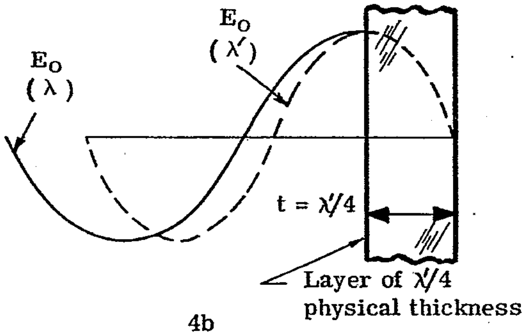


Figure 20.3 - Nomenclature used in designating the thickness, refractive index, and angle of refraction θ in each of the layers. For sake of clarity, the reflections which take place at each interface are not shown.



4a



4b

Figure 20.4 (a) - A comparison of the wavelength of light (in vacuo) with the optical thickness of a quarter-wave layer.

Figure 20.4 (b) - An instantaneous picture of the electric vector of a wave entering a denser medium. (see text 20.1.3.4) is usually quite "flat," whereas the reflectivity of a typical multilayer mirror changes quite rapidly with wavelength. Further information is presented in Section 20.8.

20.1.2.7 Heat control filters. One type of a multilayer mirror, called a cold mirror, is used to reflect the visible light and transmit the infrared. Figure 20.2 shows a typical use of a cold mirror as a reflector behind a light source in a film projector. The cold mirror reflects the visible light towards the film, but permits the heat in the infrared to pass out the back. This system is even more efficient if a heat reflector is inserted between the arc and the film. This reflector, which has a high transmission in the visible spectral region but a high reflectivity in the infrared, reflects the heat away from the film but permits the visible light to pass through to the film with little attenuation. Another type of heat control filter is the cover glass which is placed over solar cells which are used to power a space vehicle. These cover glasses are designed to reflect the radiant energy at wavelengths longer than 1.2 μ. The radiant energy from the sun in the wavelength range from 1.2 μ to 2.5 μ does not generate power, but only increases the temperature of the solar cell, thereby decreasing its efficiency. Heat control filters are discussed briefly in Sections 20.5.3 and 20.6.2.

20.1.2.8 High-reflectivity mirrors. By overcoating aluminum and other metals with dielectric films, it is possible to obtain a reflectivity as high as 99.5%. The spectral reflectivity of a typical overcoated mirror is shown in Section 20.8.

20.1.2.9 Polarizers. Multilayers can be used to produce linearly polarized light. They are particularly useful in the infrared, where conventional polarizers which utilize birefringence cannot be used because most optical materials are optically isotropic. This application of multilayers is not discussed in Section 20; the reader can refer to Heavens[1] for further details.

20.1.2.10 Reflection filters. A multilayer has been developed which has a high reflectivity in certain spectral regions, but absorbs strongly in other regions. Such a mirror has been used to absorb the visible light and reflect the infrared. It should be noted that the type of reflection filter we are discussing here is different from the heat reflector described in 20.1.2.7. Both types of multilayer filter reflect the infrared, but the former absorbs the visible light, whereas the latter transmits the visible. This type of filter is not discussed in Section 20; further information is given in Heavens[2].

20.1.3 Nomenclature

20.1.3.1 General considerations. In Section 20, we consider only the case where the thin films are optically homogeneous; that is, the optical constants of a given layer do not vary along the direction of the propagation of the light, which is shown as the Z-direction in Figure 20.3. At the present time, almost all commercially manufactured multilayer filters are composed of films which are homogeneous. However, multilayer coatings which contain optically inhomogeneous films have some very interesting properties and it is possible that they will come into more extensive use in the future. A simple antireflection coating which contains an optically inhomogeneous film, sometimes called a "graded film", is described by Strong[3].

20.1.3.2 The multilayer stack. An idealized multilayer stack is shown in Figure 20.3. It consists of a total of ℓ layers deposited upon a substrate which has an optical constant $\mu_s = n_s - j k_s$. Each of the layers has a physical thickness t_i and optical constant $\mu_i = n_i - j k_i$, where n_i is the refractive index and k_i is the absorption coefficient. It should be noted that the absorption coefficient (represented by a lower case "k") is related to the parameter "K" in Section 21.2 by the relationship $k_i = - n_\nu K_\nu$. The light is incident at an angle ϕ from a non-absorbing incident medium of refractive index n_0. Since the layer boundaries are parallel, the angle of refraction in the i^{th} layer, θ_i, is determined simply from Snell's law, if the layers are non-absorbing. Thus each layer in the stack is specified by three parameters, namely t_i, n_i, and k_i. These quantities, along with n_o, n_s, and k_s, completely specify the optical properties of the multilayer. Given these quantities and ϕ it is a straightforward, although tedious, task to calculate the reflectivity R and transmission T of a multilayer as a function of the wavelength λ of the incident light. Methods of computing R and T are presented in Sections 20.1.5, 21.2.8, and 21.2.12.

20.1.3.3 The wave number. The retardation of phase of the i^{th} layer is defined* as

$$\delta_i = 2\pi \sigma n_i t_i \qquad (1)$$

where σ is the wave number of the incident light, $\sigma = 1/\lambda$. The wave number is proportional to the frequency of the light. In computing the spectral transmission of a multilayer filter, there are many advantages to using some parameter, such as σ, which is proportional to frequency rather than wavelength. One advantage is that many spectral transmission curves have even symmetry about some point, when plotted on a frequency scale, whereas the curves are quite asymmetrical when plotted versus wavelength. This is illustrated by Figures 20.32 and 20.34, which show the reflectivity of the same coatings. It is evident that the curves plotted versus frequency, as in Figure 20.34, have even symmetry about the center, whereas the wavelength plot of Figure 20.32 is quite asymmetrical. Another advantage of using frequency as a variable is that very often maxima and minima in the reflectivity curve are spaced at equal intervals on a frequency scale, whereas on a wavelength scale they are spread out in the long wavelength region and compressed together in the short wavelength region. This is illustrated in reflectivity curve shown in Figure 20.59. The maxima and minima near 500 mμ are so close together that they cannot be plotted accurately, whereas they would be spaced at equal intervals if plotted versus σ. In many cases in Section 20 we use as a variable a dimensionless quantity, $g = \lambda_0/\lambda$, which is proportional to frequency. The disadvantage of using σ as a variable rather than λ lies in our educational system; most persons are unfamiliar with wave number. Most of us have been educated to think in terms of wavelength and thus we associate "green light" with a wavelength of 540 mμ rather than with a wave number of 18519 cm^{-1}. A useful factor to remember in converting wavelength into wave number is that

---

* S_i is related to the parameter B_ν defined in Equation 21 - (47) by the relationship $2\delta_i = \beta_\nu$.

(1) All references are listed separately at the end of this section.

$$1\mu = 1000\ \mathrm{m}\mu = 10{,}000\ \mathrm{A} \quad \text{is equivalent to } 10{,}000\ \mathrm{cm}^{-1}.$$

Thus using the fact that "ten thousand angstroms 'equals' ten thousand wave numbers", and the fact that the wave number is inversely proportional to wavelength, we see that

5000A is equivalent to 20,000 cm^{-1}

2500A is equivalent to 40,000 cm^{-1}

and so on.

20.1.3.4 The QWOT. Not infrequently multilayers are composed entirely of dielectric (non-absorbing) materials. In this case it is convenient to refer to the thickness of the layers in terms of their optical thickness, which is defined as the product of the geometrical thickness, t_i, and the refractive index, n_i. Reference is frequently made to the quarter-wave optical thickness, QWOT, which is defined as

$$\mathrm{QWOT} = 4\, n_i\, t_i\ . \qquad (2)$$

Since n_i is a dimensionless quantity and t_i has the dimensions of length, the QWOT also has the dimensions of length and is usually expressed in units of microns or millimicrons. Thus, if a layer has a QWOT of 550 mμ, this means that one-quarter wavelength of light at 550 mμ has the same length as the optical thickness of the layer, as illustrated in Figure 20.4 (a). Sometimes one refers to a film of quarter-wave optical thickness simply as a "quarter wave". A note is interpolated to delineate clearly exactly what is drawn in diagrams such as in Figures 20.4, 20.55, 20.107, and others. This represents a comparison of the optical thickness of the film with the wavelength in vacuo of the incident light. To elucidate this point, at a given instant of time, the electric vector E of a light wave propagating in a homogeneous film of refractive index n can be represented by

$$E = E_0 \cos\left(\frac{2\pi}{\lambda'} z\right)$$

where z is the coordinate shown in Figure 20.3. The wavelength λ' in the foregoing equation has a "primed" superscript as a reminder that λ' is the wavelength in that medium, of index n. λ' is related to the wavelength λ in vacuum by

$$\lambda' = \frac{\lambda}{n}$$

and consequently λ' changes when the wave enters a medium of different refractive index, as is shown in Figure 20.4b. Thus Figure 20.4b represents the actual E field in the film and vacuum at a given instant of time and the thickness of the film is represented by its actual physical thickness. Another approach, which is extensively used in Section 20, is to compare the optical thickness of the film with the wavelength in vacuo. This is accomplished by writing equation as

$$E = E_0 \cos\left(\frac{2\pi}{\lambda} n z\right).$$

The quantity nz is the optical thickness. The film shown in Figure 20.4a is drawn thicker in proportional to its optical thickness, but this is compensated for by the fact that the wavelength which is shown is not the wavelength in the film, but the vacuum wavelength.

20.1.3.5 "H" and "L" layers. Quite often multilayers are composed of films which are all of quarter-wave optical thickness, or some multiple thereof. In this case it is convenient to use a shorthand notation to specify the design. The letters "H" and "L" are used to specify films of high and low refractive index, respectively, which have the same QWOT. For example, the design of the double-layer coating shown in Figure 20.5 is designated as

glass HH L air,

where $n_H = 1.70$ and $n_L = 1.38$. The optical thickness of the film next to the glass is two quarter waves, or a single half wave. Similarly, the design of the double-layer coating shown in Figure 20.37 is

silicon H L air,

where $n_H = 2.40$ and $n_L = 1.38$.

20.1.3.6 The reflectivity and transmission. In Section 20 reference is made to computed values of the reflectivity, R, and the transmission, T. R and T are synonymous with the "reflection coefficient", and the "transmission coefficient" which are defined in terms of a time average of the Poynting vector, as is specified in Equations 45 and 45a in Section 21.2.6. In Section 20 the terms "reflectance" and "transmittance" are reserved for measured, rather than computed, values.

20.1.3.7 Non-normal incidence. When light is incident upon a multilayer at oblique incidence, both R and T must be computed separately in each plane of polarization. Thus one refers to R_p and T_p in the "p" plane of polarization when the electric vector is parallel to the plane of incidence and to R_s and T_s in the "s" plane of polarization when the electric vector is perpendicular to the plane of incidence. In general, if unpolarized light is incident upon a multilayer at non-normal incidence, both the reflected and transmitted light is partially plane-polarized. If the incident light is elliptically polarized, then the degree of elliptical polarization of both the reflected and transmitted light is altered. This is because not only the reflection and transmission are different in the two planes of polarization, but also because the phase shift upon reflection is different for the two planes of polarization. If the light which is obliquely incident upon a multilayer is initially unpolarized and if the light detector, such as the human eye or a photographic plate, is not sensitive to the polarization of the light, then the polarizing effect of the multilayer can be neglected. In this case we simply refer to the average reflectivity, R_{av}, and the average transmission, T_{av}.

$$R_{av} = \frac{1}{2}(R_p + R_s), \qquad T_{av} = \frac{1}{2}(T_p + T_s). \qquad (3)$$

20.1.4 Analogies.

20.1.4.1 Electrical transmission lines. It is useful to note the interesting analogy between the propagation of light through a thin film and the propagation of radio waves in an electrical transmission line. The propagation equation for the thin film is identical with the transmission line equation if one identifies the electric vector and magnetic vector in the thin film with the electrical voltage and current in the transmission line. The optical thickness* of the thin film is analogous to the "electrical length" of the transmission line, while the refractive index of the thin film is analogous to the "characteristic admittance" of a section of transmission line. A dielectric thin film corresponds to a "lossless" transmission line. The refractive index of the substrate and incident medium are analogous to the load admittance and the "characteristic admittance of the generator". The equations for the "optical admittance" of a thin film are identical to the equations for admittance of a transmission line. Graphical devices which have been invented for computing the voltage standing ratio of an electrical transmission line, such as the Smith Chart and the Admittance Chart, are used for computing the reflectivity of a stack of thin films[4, 5, 6]. This analogy cannot be extended too far, however. Shunt transmissions lines and lumped constant elements such as resistors, capacitors, and inductors can be added to an electrical transmission line. No exact analogous devices exist in thin film optics, although at certain wavelengths a thin gold or silver film can be represented as an inductance and an inconel film as a pure resistance.

20.1.4.2 Similarities with geometrical optics. There are some similarities between problems in multilayer filters and lens design. In both cases, it is a straightforward task to compute the performance of a given system, once the design has been specified. The design of a multilayer filters is given by specifying the thickness and optical constants of each layer, and the substrate and incident medium. The design of a lens is specified by the physical dimensions of each optical component and its refractive index. In both cases, it is quite difficult to synthesize a design. In order to synthesize a multilayer filter design, it is

---

* Note that the electrical length of a transmission line does not change with admittance as the optical thickness of a film changes with index.

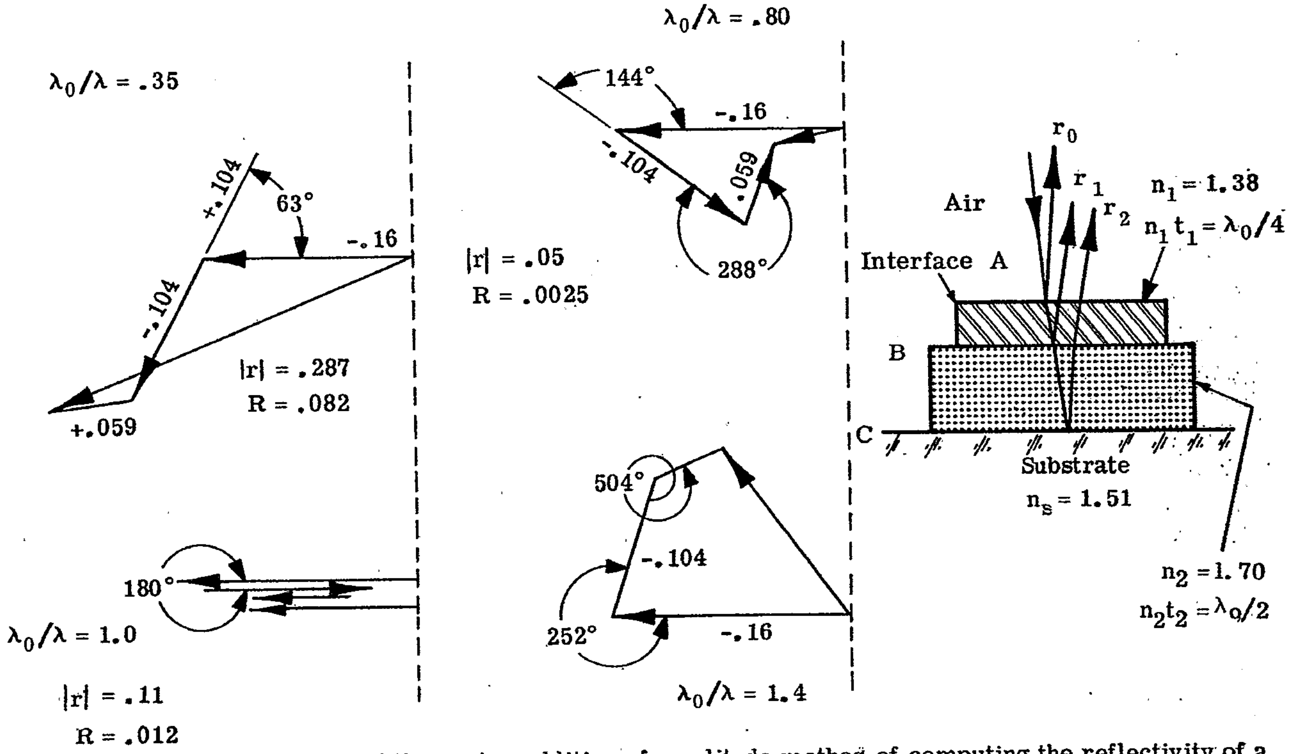


Figure 20.5-An illustration of the vector addition of amplitude method of computing the reflectivity of a antireflection coating. (Vectors are not drawn exactly to scale.)

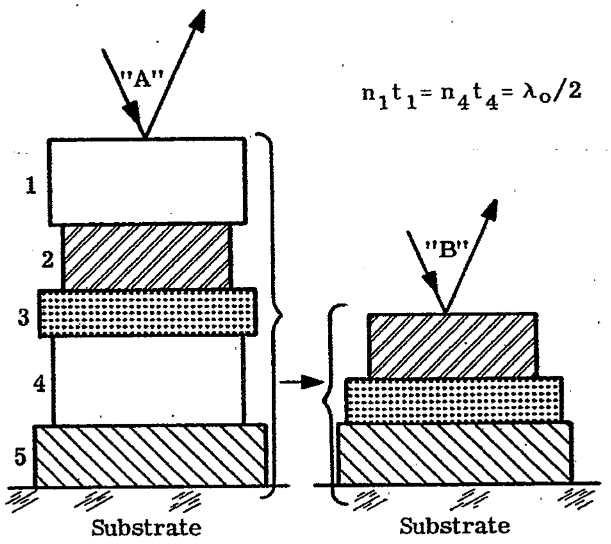


Figure 20.6-Stack "B" is equivalent to stack "A" when layers 1 and 4 in stack "A" are absentee.

necessary to choose the thickness and optical constants of each layer so that the filter has a prescribed spectral transmission or reflectivity. In synthesizing a lens, one must choose the curvature of the surfaces, spacing and refractive index of the optical elements so that the lens has a prescribed amount of aberration at various points in the image plane. In both lens design and multilayer filters, only approximate methods have been developed to accomplish this synthesis.

20.1.4.2.2 There are several differences between multilayer filters and lenses. The first difference is that the parameters which specify the design of a lens are given to a high degree of precision. The physical dimensions and the refractive indices of the optical components are usually specified to an accuracy of a few thousandths of a percent. In thin film optics, however, it is usually difficult to control the thickness of individual layers to a precision of better than one percent. The refractive index of thin films can vary widely, because the optical constants of thin film materials depend not only upon the thickness of the layer, but also quite markedly upon the conditions in the vacuum when the film is evaporated. For example, the refractive index of the material "silicon monoxide" can vary from 1.40 to 1.90 depending upon the partial pressure of oxygen during the evaporation.[7,8]

20.1.4.2.3 Another difference is that the design of a lens is usually patent, whereas the design of a multilayer can be kept secret quite easily. Even if the design of a lens were not specified, one could disassemble a lens and measure the refractive index and dimensions of its components. In order to ascertain the design of a multilayer filter, however, it would be necessary to "peel off" each of the layers, which might be as thin as a few hundred angstroms. This is practically impossible to accomplish. The amount of material in a layer which has a QWOT of 500 mμ is about thirty micrograms per square centimeter. Hence, it is quite difficult to perform a chemical analysis to determine the composition of the layers. Because of these facts, one finds quite frequently that the designs of multilayer filters and the materials which are used in their manufacture are kept secret for proprietary reasons. This secrecy has been a detriment to the progress in this field.

20.1.5 Methods of computing R and T*

20.1.5.1 Vector addition of amplitudes. This is an approximate method which is most useful for stacks which contain a small number of layers. It essentially neglects the multiple reflections which take place between various interfaces and hence is most accurate when the difference between the refractive indices of adjacent layers is small. As an example of the application of this method, consider the problem of computing reflectivity of the double-layer antireflection coating shown in Figure 20.5 . The incident light reflects from each of the three interfaces. The amplitude of the wave which is reflected at each interface is proportional to the Fresnel amplitude coefficient

$$r_i = (n_i - n_{i+1})(n_i + n_{i+1})^{-1} \tag{4}$$

where n_{i+1} and n_i refer to the refractive index on each side of the interface. However, these waves are not in phase and hence their amplitudes must be added vectorially. The difference in phase between the wave reflected from interfaces A and B is $2\,\delta_1$, and similarly for the other interface. In the example in Figure 20.5 , the r_i are -0.16, -0.104, and 0.059, respectively for interfaces A, B, and C. From Equation 20-(1) we readily determine that 2_1 is 63° and 2_2 is 126° when $\lambda_o/\lambda = 0.35$. Extending this procedure to other values of λ_o/λ , we can construct the vector diagrams which are shown in Figure 20.5 . When $\lambda_o/\lambda = 1.0$, the vectors are colinear and hence the r_i can be added algebretically. In Figure 21.8 the reflectivity computed by the vector method is compared with the results of the more exact matrix method described in the next section.

20.1.5.2 The characteristic matrix.[11,12]

20.1.5.2.1 The electric field E and the magnetic field H at one boundary of a film are related to the fields E' and H' at the other boundary by two linear simultaneous algebretic equations. These equations can be written in matrix form:

$$\begin{bmatrix} E \\ H \end{bmatrix} = M_i \begin{bmatrix} E' \\ H' \end{bmatrix} \tag{5}$$

where the matrix M for a non-absorbing film at normal incidence is

$$M_i = \begin{bmatrix} \cos \delta_i & j\, n^{-1}_i \sin \delta_i \\ j\, n_i \sin \delta_i & \cos \delta_i \end{bmatrix} \tag{6}$$

---

* The derivations of the equations which are cited in this section for computing R and T are given in section 21.4 or in references 9 and 10.

and where $j = \sqrt{-1}$ and $\begin{bmatrix} E \\ H \end{bmatrix}$ and $\begin{bmatrix} E' \\ H' \end{bmatrix}$ represent column vectors. As is shown in Section 21.4, Equation 20-(5) can readily be extended to more than one layer and thus this can be regarded as a recursion relationship. The reflectivity of a multilayer is computed by first writing down the matrix M_i for each layer according to Equation 20-(6). Then the matrix product is computed, as for example in Equation 21-(129) in Section 21.4. For example, if a stack consists of three layers, the matrix product M is

$$M = M_1 * M_2 * M_3 \tag{7}$$

where the symbol $*$ denotes a matrix multiplication. The matrix product M has in general four elements:

$$M = \begin{bmatrix} A & j\,B \\ j\,C & D \end{bmatrix} . \tag{8}$$

The four variables, A, B, C, and D, are all real variables if all the films are non-absorbing. However, only three of the variables are independent, since the determinate of the matrix M is unity and hence $A \bullet D + B \bullet C = 1$. The reflectivity R is computed from

$$R = \frac{(X - U)^2 + (Y - V)^2}{(X + U)^2 + (Y + V)^2} \tag{9}$$

where

$$X = n_o\, A + n_o\, k_s\, B \qquad U = n_s\, D$$
$$Y = n_o\, n_s\, B \qquad V = C - k_s\, D \tag{10}$$

If the substrate is non-absorbing, i.e. $k_s = 0$, conservation of energy requires that $R + T = 1$ and in this case T can be written:

$$T = \frac{4}{2 + A^2 \dfrac{n_o}{n_s} + D^2 \dfrac{n_s}{n_o} + \dfrac{C^2}{n_o\, n_s} + B^2\, n_o\, n_s} . \tag{11}$$

20.1.5.2.2 Whenever the optical thickness of any layer is $\lambda/2$, λ, $3\lambda/2$, 2λ, etc. the δ of that layer is 180°, 360°, 540°, etc. and the characteristic matrix for that layer reduces to the unit matrix:

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} . \tag{12}$$

However, the unit matrix has no effect upon the matrix product and hence this "half-wave" film does not contribute to the reflectivity of the multilayer stack at that wavelength. In other words, any "half-wave" layer acts as though it is absent from stack - that is, it is an absentee layer (see Section 21.2.14). As an example, consider the five-layer stack which is shown in Figure 20.6. It is specified that at some wavelength, λ_o, optical thickness of layers #1 and #4 is $\lambda_o/2$. Then at λ_o layers #1 and #4 are absentee, and the reflectivity of the five layer stack is equivalent to the reflectivity of the three-layer stack shown in Figure 20.6, which is the five-layer stack with layers #1 and #4 removed.

20.1.6 The computation of R and T at non-normal incidence.

20.1.6.1 Extension of the normal-incidence equations. Equations 20-(4), 20-(6), 20-(10) and 20-(11) strictly apply to normal incidence. However, they can be extended to include a non-absorbing multilayer stack at non-normal incidence. This is accomplished as follows:[13]

Step 1. Given the angle ϕ in the incident medium, the angle of refraction, θ_i, is computed from Snell's law.

Step 2. The "s" plane of polarization is considered first. The refractive index of the layer, n_i, as it appears explicitly in Equations 20-(4) and 20-(6) is replaced by an effective index,

$$n_{eff} = n_i \cos \theta_i . \tag{13}$$

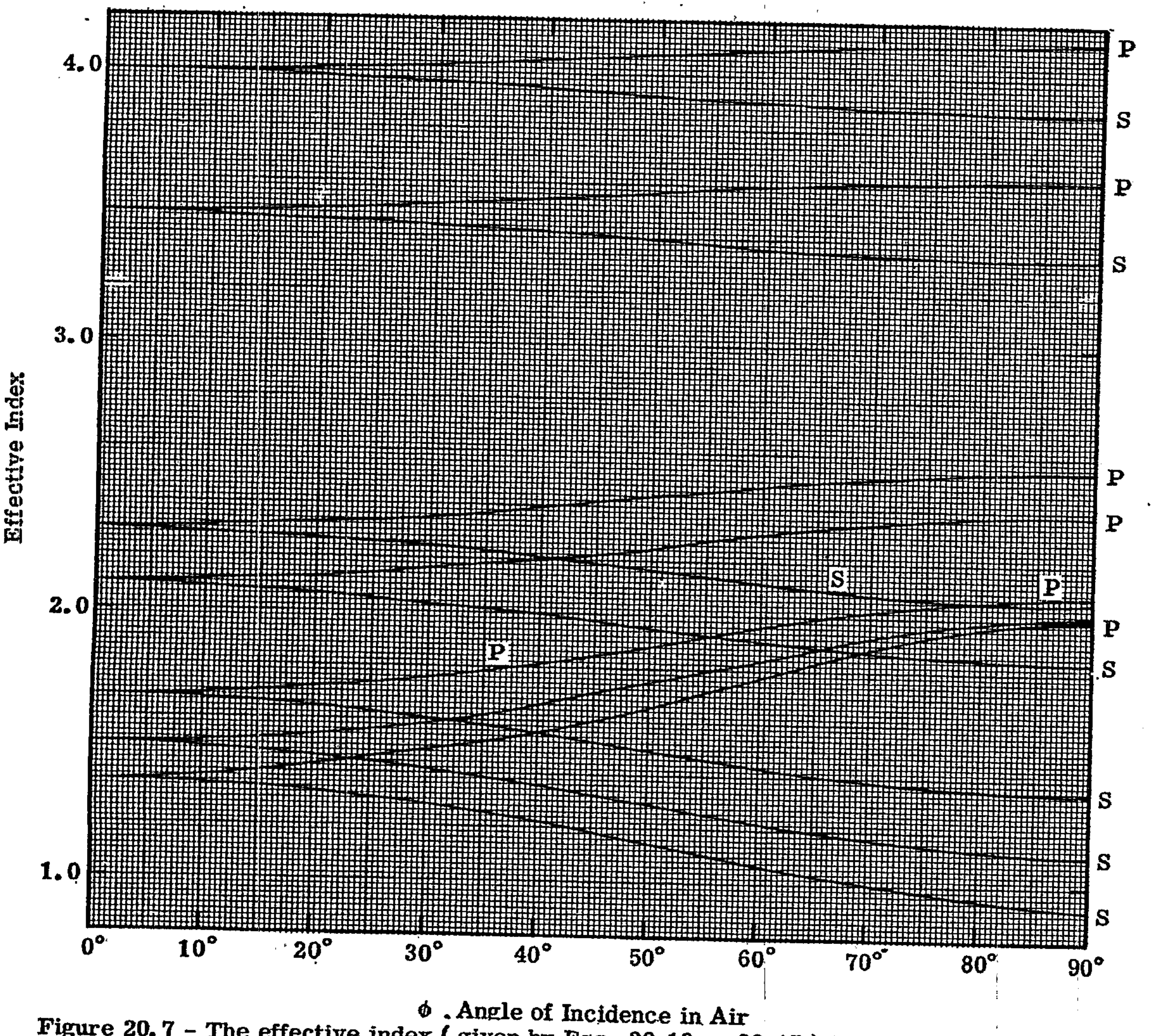


Figure 20.7 - The effective index (given by Eqs. 20.13 or 20.15) for the " p "and "s" planes of polarization, as a function of the angle of incidence in air.

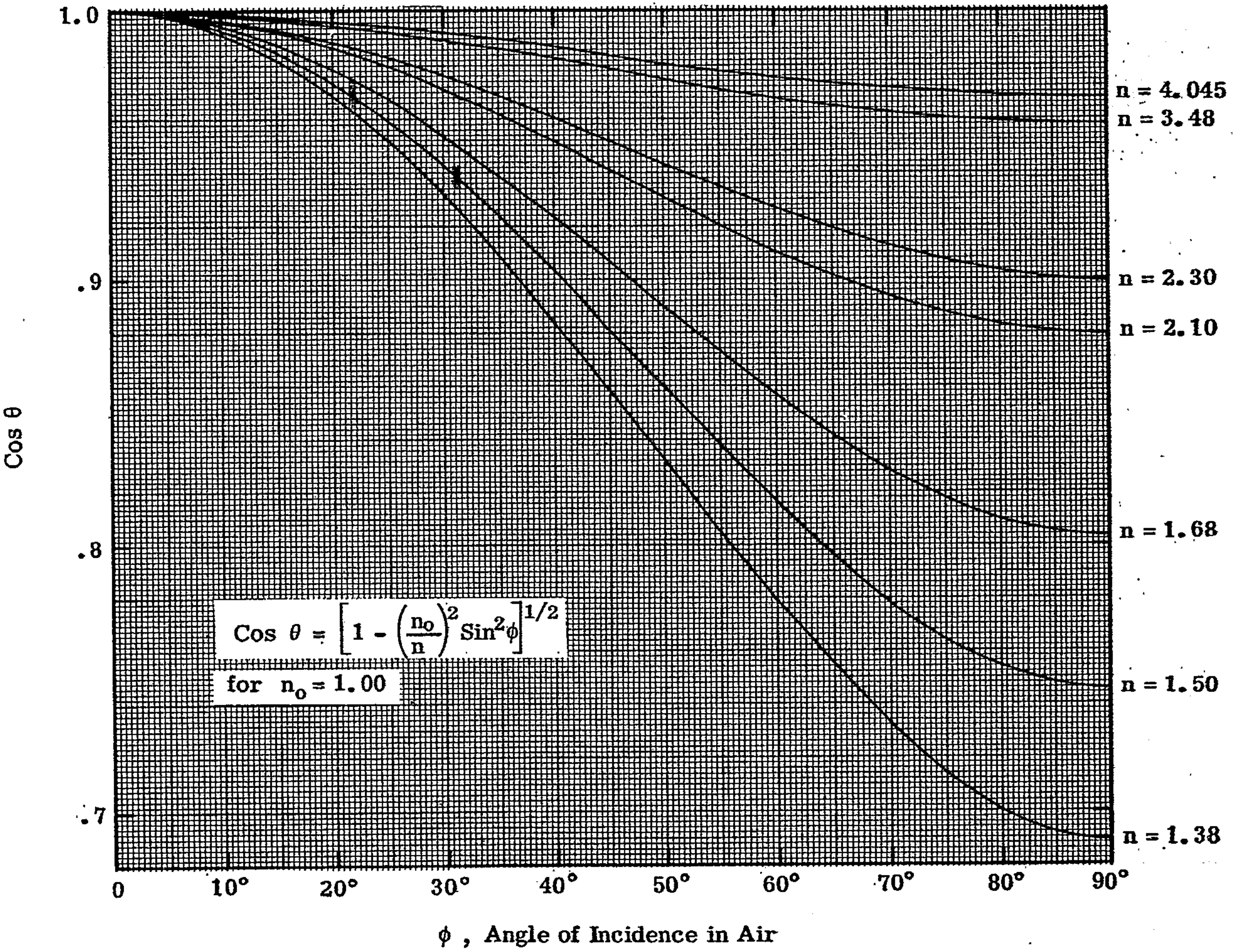


Figure 20.8 The effective thickness (normalized to 1.00 at normal incidence) as a function of the angle of incidence in air.

This substitution does not apply to the n_i which appears in Equation 20-(1). The indices n_o and n_s are replaced by the effective indices $n_o \cos \phi$ and $n_s \cos \chi$, respectively.

Step 3. The optical thickness of each layer is replaced by an effective thickness,

$$(n_i \; t_i)_{eff} = n_i \; t_i \; \cos \theta_i \; . \tag{14}$$

Step 4. Having executed steps 1, 2, and 3, R_s and T_s are computed, using either the vector addition of amplitude method or the matrix method.

Step 5. The "p" plane of polarization is considered next. Rather than Equation 20-(13), we use for the effective index:

$$n_{eff} = n_i \,/\, \cos \theta_i \; . \tag{15}$$

Again, Equation 20-(15) does not apply to the n_i in Equation 20-(1). The indices n_o and n_s are replaced by the effective indices $n_o / \cos \phi$ and $n_s / \cos \chi$, respectively.

Step 6. Having made the substitutions indicated in steps 3 and 5, R_p and T_p are computed, using either the vector method or the matrix method.

20.1.6.2 Use of the effective index.

20.1.6.2.1 Figure 20.7 shows the plot of the effective index as a function of ϕ for an incident medium of refractive index $n_o = 1.00$. As ϕ approaches 90°, the effective index approaches the limiting value of

$$n_{eff} = \sqrt{n_i^{\,2} - n_o^{\,2}} \tag{16}$$

and

$$n_{eff} = n_i^2 \,/\, \sqrt{n_i^{\,2} - n_o^2} \tag{17}$$

for the "s" and "p" planes of polarization, respectively. From Figure 20.7 we see that the effective index of a material with a large refractive index, such as germanium, changes by less than three percent between $\phi = 0$ and $\phi = 90°$. The materials with a lower index show a much larger change.

20.1.6.2.2 Figure 20.8 shows the fractional change in the effective thickness, $\cos \theta_i$, as a function of ϕ for various values of the index n_i. As one might expect, the change in the effective thickness between $\phi = 0$ and 90° is much greater for low-index materials than for high-index materials.

20.1.6.2.3 Since the effective thickness at oblique incidence is always less than at normal incidence, this means that the reflectivity and transmission peaks of multilayer filters shift to shorter wavelengths as ϕ increases. Although the author does not know of a rigorous proof of the statement made in the foregoing sentence. he has never found any exception to it in his work with multilayer filters.

20.1.6.3 Matched layers.

20.1.6.3.1 By definition, two or more layers in a multilayer stack are matched when there is a prescribed ratio between the optical thickness, $n_i \; t_i$, and hence between the δ_i of those layers.

20.1.6.3.2 As an example, consider the three-layer stack which is shown in Figure 20.9. For the purpose of this illustration, we shall arbitrarily define a matched condition to be:

$$n_1 \; t_1 \; : \; n_2 \; t_2 \; : \; n_3 \; t_3 = 1 : 2 : 2.5 \, . \tag{18}$$

Equation 20-(18) states that when the optical thickness of the second layer is twice that of the first layer and the $n_3 \; t_3$ of layer three is 2.5 times $n_1 \; t_1$. If this film combination is tipped at an angle of 60°, then the optical thickness of each layer must be replaced by its effective thickness. Since the refractive index for each layer is different, the percentage change in the effective thickness is different. Referring to Figure 20.8 we see that the optical thickness of layer one is multiplied 0.98, layer two by 0.78, and layer three by 0.86.

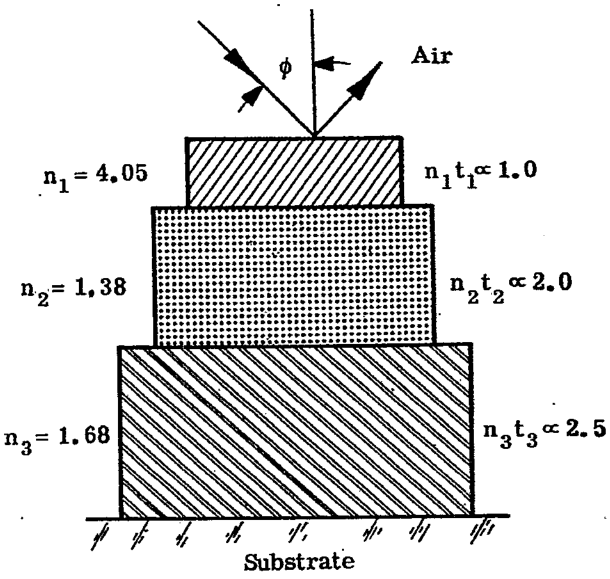


Figure 20.9 - Diagram of a three-layer stack. The layers are drawn proportional to their optical thickness.

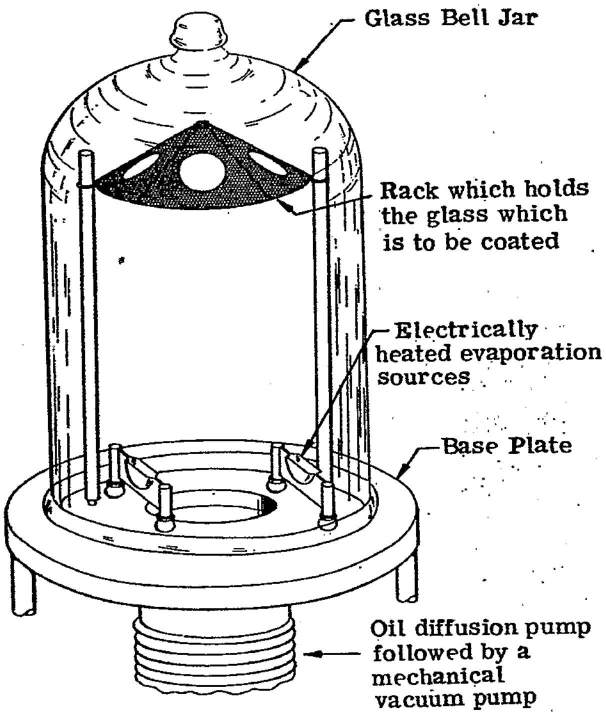

Glass Bell Jar
Rack which holds the glass which is to be coated
Electrically heated evaporation sources
Base Plate
Oil diffusion pump followed by a mechanical vacuum pump

Figure 20.10 - The essential parts of an evaporator which is used to deposit thin films.

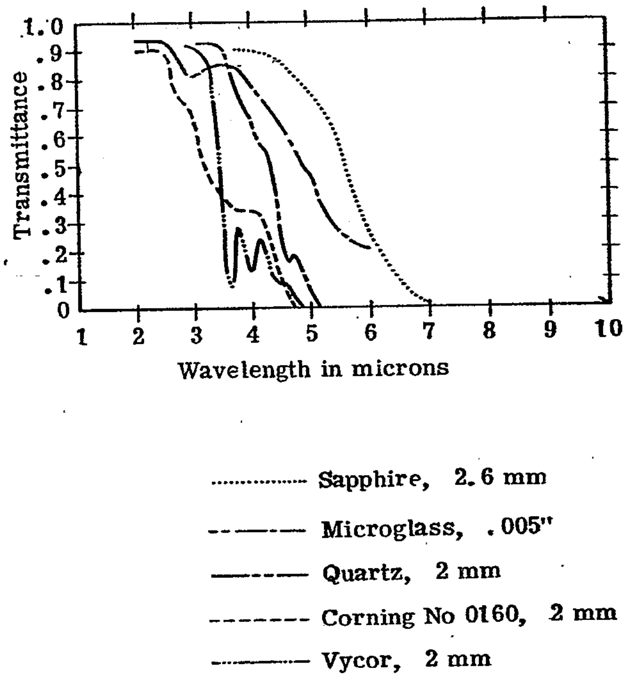


Figure 20.11 - The optical transmission in the infrared of some substrate materials.

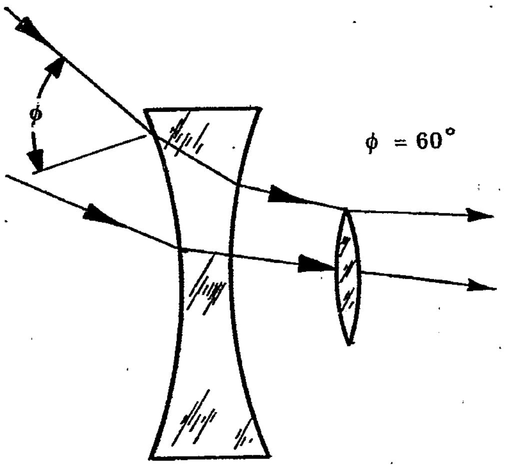


Figure 20.12 - The light is incident at a large angle upon the front surface of a wide-angle lens.

This means that at 60° incidence, the ratio of the thickness of the layers is no longer 1 : 2 : 2.5, but is

$$\left[n_1 \ t_1 \right] : \left[n_2 \ t_2 \right] : \left[n_3 \ t_3 \right] = 0.977 : 1.56 : 2.14 = 1 : 1.60 : 2.19 \qquad (19)$$

where the brackets $[\ \ ]$ indicate "effective thickness". This means that the layers are no longer matched at 60° incidence, or for any other non-zero angle, for that matter. If all of the layers were composed of high refractive index materials, the deviation of the optical thickness ratios from the matched condition prescribed in Equation 20-(18) would be much smaller.

20.1.6.4 Layers matched at oblique incidence.

20.1.6.4.1 From the discussion in the foregoing subsection, we see that if the layers were deliberately made thicker in proportion to sec θ_i then the layers are matched at non-normal incidence for a particular value of ϕ. In the foregoing example, if the $n_i \ t_i$ of the layers were in the ratio of

$$n_1 \ t_1 : n_2 \ t_2 : n_3 \ t_3 = 1.024 : 2.57 : 2.92 \qquad (20)$$

then at an incidence angle of 60° the layers will be matched with the prescribed ratio stated in Equation 20-(18).

20.1.6.4.2 It only makes sense to speak of matched layers provided there is more than one layer. In the case of a single layer if the $\delta_1 = 2\pi/\lambda \ (n_1 \ t_1)$ of that layer has some value at normal incidence, say, for example, 1.25 radians at some wavelength, λ_1. Then as ϕ increases, the effective thickness decreases. It is always possible to find some new wavelength λ_2, at which δ_1, is still 1.25 radians. If $n_o < n_1$, then λ_2 is less than λ_1. This shift to shorter wavelengths is clearly illustrated in reflectivity curves of single-layer reflecting coatings which are shown in Figure 20.20.

20.2 THE MANUFACTURE OF MULTILAYER FILTERS

20.2.1 Practical considerations. Using the computational methods which were described in 20.1.5, the spectral transmission and reflectivity of a given multilayer filter can be computed. However, if this filter is not to be a mere theoretical abstraction, but is to be actually manufactured, then we must keep in mind that there are certain practical problems which are encountered. Just as the lens designer is limited by the fact that the optical glasses which are presently manufactured have refractive indices within in a certain range, the films of a multilayer filter must be composed of materials which have certain specific values of refractive index.

20.2.2 Methods of deposition.

20.2.2.1 Progress in optical filming. Multilayer filters are manufactured by depositing solid films of various materials on an appropriate substrate. Although there are many methods which are used to deposit these films, such as chemical reaction in a vapor or liquid, sputtering, or anodization, the most important and widely-used methods is evaporation in a vacuum.* Consequently, progress in the production and manufacture of multilayer filters has for many years been related to advances in vacuum technology. Although the theory of multilayer filters has been known for more than a century, the production of such useful devices as antireflection coatings for lenses and silver-dielectric-silver interference filters did not start until the decade of the 1930's, when high-capacity oil diffusion pumps were developed which could evacuate a large volume to a pressure of less than 10^{-7} of atmospheric pressure. Antireflection coatings for lenses were used extensively during the Second World War. After the war, all-dielectric interference filters and multilayer beam splitters for use in the visible spectral region became commercially available. In the decade from 1950 to 1960, multilayer coatings have been produced for wavelengths as short as 110 $m\mu$ in the ultraviolet[15] and for wavelengths as long as 20μ in the infrared.[16]

20.2.2.2 The evaporator. Figure 20.10 shows the essential parts of an evaporator which is used to deposit thin films by evaporation in a vacuum. The circular pieces which are to be coated (i.e. the substrates) are placed in the holes in holder at the top of the chamber. The glass bell jar is sealed to the base plate and the entire chamber is evacuated by means of the oil diffusion pump and the mechanical pump to a pressure of less than 10^{-4} mm of mercury. The boats which contain the material which is to be evaporated are then electrically heated causing the material in the boat to evaporate and deposit in a thin solid film on the substrates. No mention has been made here of how the substrates are cleaned so that the films adhere well, or how the films are evaporated to a predetermined thickness. These topics are covered in detail in references 17 and 18.

---

* Methods of depositing thin films are summarized in reference 17.

MATERIAL	REFRACTIVE INDEX	TRANSMISSION RANGE .	COMMENTS
Hot pressed Mg F_2 (Irtran 1)	1.37	200 mμ to 7.5 μ	
Calcium fluoride	1.42	180 mμ to 6 μ	
Barium fluoride	1.42	150 mμ to 13 μ	
Irtran 3	1.42	1000 mμ to 9 μ	
Fused quartz	1.46	180 mμ to 3.5 μ	1
Vycor (high-silica glass)	1.46	250 mμ to 3.5 μ	
Glass	1.51 to 1.70	320 mμ to 2.5 μ	2
Sapphire	1.70	<20 mμ to 7.5 μ	
Hot pressed ZnS (Irtran 2)	2.26	2000 mμ to 14 μ	
Arsenic trisulfide glass	2.40	800 mμ to 12 μ	
Irtran 4	2.40	1000 mμ to 20 μ	
Silicon	3.50	1100 mμ to 8 μ	
Germanium	4.05	1900 mμ to >22 μ	

NOTES:

1. The wavelength of the absorption edge in the ultraviolet depends upon the purity of the quartz.
2. In general, glasses with a lower refractive index are transparent to shorter wavelengths than glasses with a higher index. However, there are exceptions to this rule. Most glasses have a strong absorption band in the vicinity of 2.6 μ. Very thin glass plates (i.e. "cover slips") are transparent between 2.7 μ and 4.5 μ.

Table 20.1 - Common substrate materials.

MATERIAL	REFRACTIVE INDEX	RANGE OF TRANSPARENCY (See Note 10)		COMMENTS
		From	To	
Cryolite	1.35	<200 mμ	10 μ	1
Chiolite	1.35	<200 mμ	10 μ	1
Magnesium fluoride	1.38	230 mμ	5 μ	2, 3
Thorium fluoride	1.45	<200 mμ	10 μ	
Cerium fluoride	1.62	300 mμ	> 5 μ	4
Silicon monoxide	1.45 to 1.90	350 mμ	8 μ	5
Sodium chloride	1.54	180 mμ	>15 μ	6
Zirconium dioxide	2.10	300 mμ	> 7 μ	2
Zinc sulfide	2.30	400 mμ	14 μ	7
Titanium dioxide	2.40 to 2.90	400 mμ	> 7 μ	8
Cerium dioxide	2.30	400 mμ	5 μ	2, 3
Silicon	3.50	900 mμ	8 μ	
Germanium	3.80 to 4.20	1400 mμ	> 20 μ	9
Lead telluride	5.10	3900 mμ	> 20 μ	

NOTES:

1. Both materials are sodium-aluminum fluroide compounds, but differ in the ratio of Na to Al and have different crystal structure. Chiolite is preferable in the infrared, because it has less stress than cryolite (see section 20.2.3.2.4).
2. These materials are hard and durable, especially when evaporated onto a hot substrate.
3. The long wavelength is limited by the fact that when the optical thickness of the film is a quarter-wave at 5 μ, the film cracks due to the mechanical stress (see 20.2.3.2.4).
4. There are other fluorides and oxides of rare earths which have refractive indices in this range from 1.60 to 2.0. See reference 22a.
5. The refractive index of SiO_x (called silicon monoxide) can vary from 1.45 to 1.90 depending upon the partial pressure of oxygen during the evaporation. Films with a refractive index of 1.75 and higher absorb at wavelengths below 500 mμ.
6. Sodium chloride is used in interference filters out to a wavelength of 20 μ. It has very little stress.
7. The refractive index of zinc sulfide is dispersive.[23]
8. The refractive index of TiO_2 rises sharply in the blue spectral region.[24]
9. The higher refractive index of 4.2 is given in reference 23. The lower index is quoted by Dr. A. F. Turner of Rochester, N.Y. (private communication).
10. The range of transparency is for a film of quarter-wave optical thickness at this wavelength. These values are approximate and also depend quite markedly upon the conditions in the vacuum during the evaporation of the film.

Table 20.2 - Commonly used thin film materials.

20.2.3 Substrates for multilayer filters.

20.2.3.1 Optical characteristics. If transmission-type filters are used, the substrate must not absorb in the spectral transmission range of the filter. Very often the fact that a substrate material absorbs in certain wavelength regions can be used to advantage. For example, suppose a filter is required for the infrared which has a high transmission at wavelengths longer than 7.0 μ , and a high attenuation at shorter wavelengths. If a germanium substrate were used, then the germanium would absorb at wavelengths shorter than 1.8 μ . Thus the multilayer which is deposited on the germanium would be required to attenuate between 1.8 μ and 7.0 μ . On the other hand, if a substrate of silicon were used, which absorbs at wavelengths below 1 μ , then the multilayer coatings would have to attenuate over a wider range of wavelengths and hence the multilayer coatings would be more complex and expensive to fabricate. The infrared transmission of some optical materials which can be used as multilayer substrates is shown in Figure 20.11. Ballard, McCarthy,and Wolfe [19] have published transmission data on other materials. The refractive index of the substrate is also important. A high-index substrate has high reflection losses at its surfaces if it is not properly coated with antireflection coatings.

20.2.3.2 Chemical and physical properties.

20.2.3.2.1 Certain substrate materials, such as sapphire and fused quartz are hard, durable, and relatively inert to chemical attack. Other materials, notably the rare-earth glasses and some of the alkali halide crystals, are rather soft and delicate. Multilayer coatings are deposited on a substrate and then sometimes rejected because their optical transmission does not meet specifications. If the substrate is expensive, it is desirable to remove the coating from the substrate and recoat it. It is advantageous to use a substrate which is chemically inert, because in this case the coating can be removed with acid or alkali solutions. Otherwise, it is necessary to remove the coating mechanically by the more expensive method of optical polishing.

20.2.3.2.2 The adhesion of the coatings to the substrate is also important. For example, sapphire is a hard and durable substrate, but it has the disadvantage that some thin film materials do not adhere well to its surface.

20.2.3.2.3 The fragility of the substrate is also a consideration. Some thin film materials, such as magnesium fluoride and cerium dioxide, are much harder and more durable when they are evaporated onto a substrate which is heated to a relatively high temperature, often as high as 300° C. It is not an easy task to heat a large piece of optical glass to this temperature without fracturing it, whereas a fused quartz substrate could easily withstand this heating.

20.2.3.2.4 The number of available substrate materials is legion and an exhaustive list would be quite long. Some of the commonly used materials are listed in Table 20.1.

20.2.4 Thin film materials.

20.2.4.1 Optical properties. In most cases, the thickness and refractive index of the films in a multilayer filter are chosen from theoretical considerations. In order to translate this design into a practical filter, it is necessary to select for each layer a thin film material which can be evaporated to the desired thickness and which has a refractive index which is close to the theoretical value. We see from Table 20.2 that in the visible spectral region, non-absorbing films are available with a refractive index in the range from 1.35 to 2.70. In the infrared, materials such as silicon, germanium, and lead telluride are available which have a considerably higher refractive index. In most cases the film does not have the same refractive index as the bulk material. A comprehensive list of thin film coating materials is compiled by Heavens. [22]

20.2.4.2 Physical properties.

20.2.4.2.1 In some applications it is possible to protect a multilayer by cementing on top of it another transparent plate. Thus most silver-dielectric-silver FP type filters (described in 20.10.3.1) are protected in this manner by "sandwiching" the multilayer between two glass plates. If a multilayer is protected, then it is possible to use materials in the stack which are "soft", such as antimony trioxide. In other cases, the multilayer must be "hard" to resist scratching and abrasion because it is exposed and subjected to extreme environmental conditions. Such a coating is called a "hard coating" or simply a "hard coat". A commonly used means of testing the durability of a multilayer is to see if it can withstand rubbing with a soft rubber eraser. Certain coating materials, notably magnesium fluoride, silicon monoxide, cerium dioxide, titanium dioxide, germanium, and silicon, are extremely hard and durable. This does not imply that these materials are the only "hard" coatings. In fact, we must resist the temptation to classify every coating material as either "hard" or "soft" and remember that there is a continuous variation of the durability between the extremely hard materials mentioned in the foregoing sentence and a fragile material such as antimony trioxide. The durability of the coating also depends markedly on the conditions in the vacuum during the evaporation. For example, techniques have been developed to evaporate zinc sulfide so that it forms a quite durable layer. The durability of a multilayer filter depends not only on the materials, but also on the technical competence of its manufacturer.

20.2.4.2.2 Resistance to moisture. Some coating materials, such as magnesium fluoride and silicon monoxide, can be immersed in water and even in hot saline solutions for some length of time, with no deleterious effects. Other materials, such as cryolite or chiolite, have quite desirable optical properties, but are not widely used as antireflection coatings because they are slightly water soluble. Even though some coating materials are relatively insoluble, moisture can still destroy a multilayer by destroying the bond between the films and the substrate.

20.2.4.2.3 Adhesion to the substrate. It is important that the multilayer coating adhere strongly to the substrate. This is especially true of coatings which have a mechanical stress. Over the years a vast amount of lore has been acquired by the manufacturers of thin film coatings as to what film materials adhere well to various substrates. The adhesion of a film to a substrate depends quite markedly such parameters as the cleanliness of the substrate, its temperature during the evaporation, and the partial pressure of residual gases in the vacuum chamber during the evaporation. There is no substitute for experience in acquiring the "know-how" of producing durable and adherent coatings, although Holland[18] is a source of much useful information.

20.2.4.2.4 Mechanical stress.[20] Early workers found that after a multilayer filter had been evaporated and was moved from the vacuum into the humid air of the laboratory, the entire coating would separate from substrate or craze with many fine cracks. Further investigation showed that this was due to a mechanical stress in the film which is proportional to the thickness of the film.[20] This stress can be demonstrated quite easily by observing that a thin substrate bends as a film is deposited upon it. The substrate bends concave towards the evaporation source if the stress is compressive. It was found that almost all thin film materials have a tensive stress; one exception is zinc sulfide, which has a compressive stress of 0.02 (in arbitrary units). Magnesium fluoride, on the other hand, has a tensive stress of 0.11 units.

20.2.4.2.5 Limit on the thickness.[20] It is clear that this stress is one of the principal factors which limits the thickness of films which are deposited by evaporation. When the adhesion of the film to the substrate can no longer balance the stress which builds up as the film grows thicker during an evaporation, the film either parts from the substrate or crazes. Films of either magnesium fluoride or cerium dioxide tend to craze when their physical thickness exceeds one micron. Thus film materials which have a high stress are limited in their application in the infrared, where the films must be quite thick. Another factor which limits the use of some films in the infrared is called the clouding effect. This effect is not observed in a film which is relative thin, say 100 mμ in physical thickness, and hence it is transparent. However, as a much thicker film is evaporated, the film becomes cloudy and scatters light like a ground slab of glass. This is presumably due to the growth of large crystallites in the film which scatter light.

20.2.4.2.6 An effective way of avoiding stress is to use thin film materials which have little or no stress, such as silicon monoxide or sodium chloride. The latter material could only be used in the laboratory, due to its water solubility. Another approach is to deposit a film with compressive stress next to a film with a tensive stress, the thicknesses of each layer being chosen so that the total stress of the pair of films is zero.[20,21]

20.3 ANTIREFLECTION COATINGS

In this section we shall consider the problem of reducing the reflectivity of a dielectric substrate by the addition of one or more non-absorbing thin films. In order to determine how much the transmission of the substrate has been improved by the addition of the multilayer coating, we must first consider the reflectivity of the uncoated substrate.

20.3.1 Reflectivity of an uncoated surface. The reflectivity R for a bare dielectric surface at normal incidence is given by the Fresnel coefficient,

$$R = 1 - T = (n_o - n_s)^2 / (n_o + n_s)^2 \qquad (21)$$

where n_o is the refractive index of the incident medium (which is usually air) and n_s is the refractive index of the substrate. If the light is obliquely incident, then Equation 20-(21) still applies if the appropriate effective index is used in place of n_o and n_s.

20.3.2 Choice of type of coating.

20.3.2.1 Number of layers. In Section 20.3, the spectral reflectivity of antireflection coatings consisting

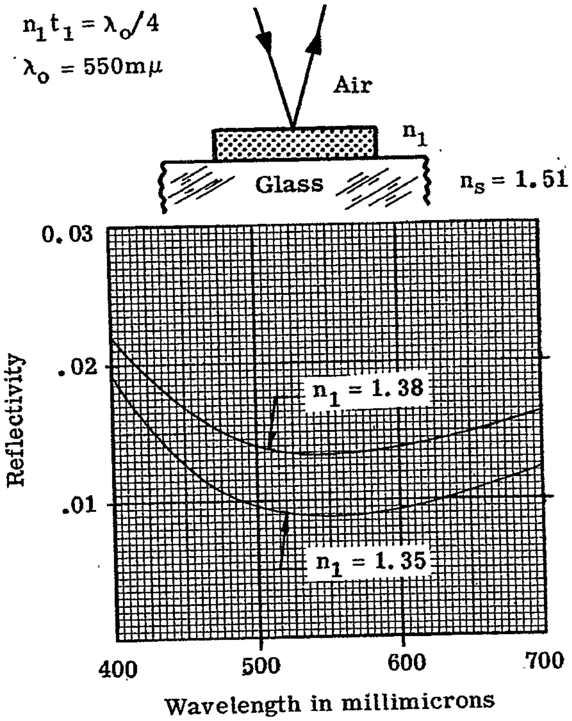


Figure 20.13 - Computed spectral reflectivity of single layers at normal incidence.

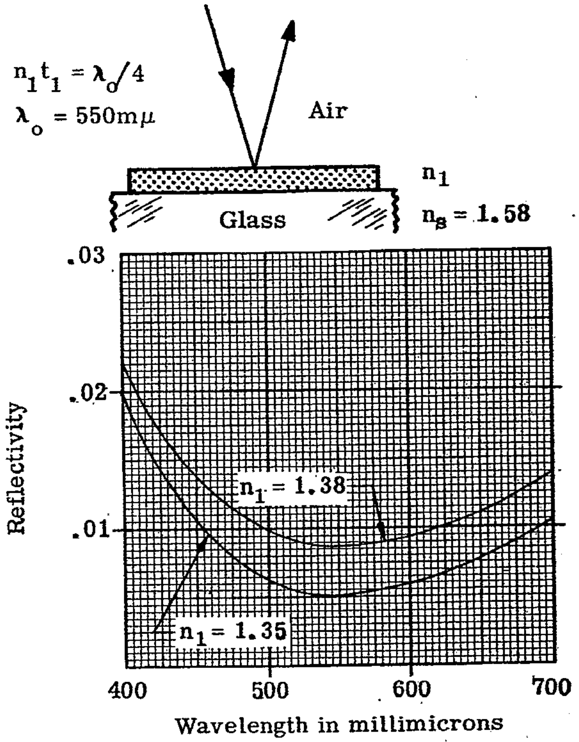


Figure 20.14 - Computed spectral reflectivity of single layers at normal incidence.

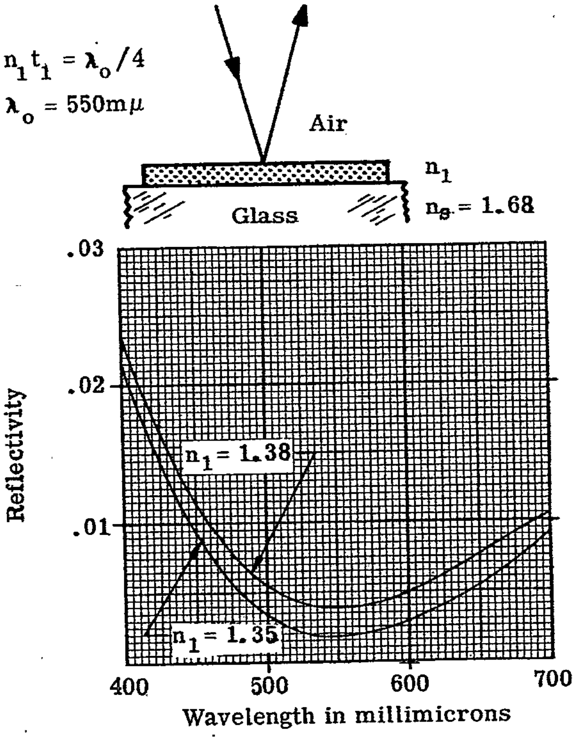


Figure 20.15 - Computed spectral reflectivity of single layers at normal incidence.

of one, two, or three layers is considered. Coatings consisting of more than three layers can be devised, but coatings with less than three layers have proved adequate for most applications; hence coatings with more layers are rarely used. In deciding what type of coating to use, the following points are usually considered:

20.3.2.2 Spectral reflectivity. The type of coating which is selected is often determined by breadth of the spectral region, over which the surface is to have a low reflectivity. For example, suppose that a lens with a large number of elements is used to image a source which emits a considerable amount of radiant energy in the narrow wavelength range from 500 mμ to 550 mμ and a negligible amount of radiant energy at other wavelengths. The best type of antireflection coating for the glass surfaces of this lens would be the two-layer coating which is described in Section 20.3.4.3.2. This coating has a very low reflectivity over a narrow wavelength region and a reflectivity which rises sharply outside of that region. On the other hand, if the source were to emit radiation over a much broader spectral region, say from 400 to 800 mμ, then other types of coatings should be considered, such as a single-layer or three-layer coating, which have a low reflectivity over a much wider wavelength region.

20.3.2.3 Angle of incidence. The reflectivity of all thin film coatings changes with the angle of incidence of the light. The type of antireflection coating which is selected might depend upon the angle of incidence of the light and the amount of convergence in the beam. For example, suppose that a coating is selected for the surfaces of the large negative lens which is the first element in a wide-angle camera lens, as shown in Figure 20.12. The angle of incidence ϕ on the first surface of this lens can easily be as high as sixty or seventy degrees. Thus if excessive vignetting is to be avoided, the antireflection coating on that surface should have a low reflectance at high angles of incidence as well as at normal incidence.

20.3.2.4 Cost. The cost of an antireflection coating is related to complexity of the equipment necessary to evaporate the layers to a prescribed thickness and also to the number of layers in the coating. For example, single-layer coatings of magnesium fluoride are quite easy to produce. The thickness of the magnesium fluoride layer can be easily determined visually during the coating process by examining the color of the reflected light. There are many facilities which have the capability of depositing these coatings because they are quite easy to produce. The production of a three-layer coating requires more elaborate equipment, such as photoelectric monitoring equipment to measure the thickness of the layers and hence the coatings are more expensive.

20.3.3 Single-layer coatings.

20.3.3.1 Basic equations.

20.3.3.1.1 The reflectivity of a dielectric surface coated with a single layer of refractive index n_1 is,

$$R = 1 - T = \frac{a_1 \cos^2 \delta_1 + a_2 \sin^2 \delta_1}{a_3 \cos^2 \delta_1 + a_4 \sin^2 \delta_1} \tag{23}$$

where

$$a_1 = (n_o - n_s)^2 \qquad a_2 = (n_1 - n_o n_s/n_1)^2$$

$$a_3 = (n_o + n_s)^2 \qquad a_4 = (n_1 + n_o n_s/n_1)^2$$

where n_o, n_s, and δ_1 have been defined previously in Section 20.1.3.2. The foregoing equation can be derived from Equations (6), (9), and (10) in Section 20.1.5.2. From Equation (23) one can see that when $n_1 t_1 = \lambda/2, 3\lambda/2, 5\lambda/2$, etc., δ_1 is 180°, 360°, 540°, etc. and hence the layer is an absentee layer (defined in 20.1.5.2.2). In this case, the reflectivity is the same as an uncoated surface and Equation (23) reduces to Equation (21).

20.3.3.1.2 When the $n_1 t_1 = \lambda/4, 3\lambda/4, 5\lambda/4$, etc., the reflectivity is either a maximum or a minimum and is given by

$$R_m = \left[\frac{n_1^2 - n_o n_s}{n_1^2 + n_o n_s}\right]^2 . \tag{24}$$

Some curves of the spectral reflectivity of a single layer on a dielectric substrate are shown in Figures 21.11, 21.12, and 21.13. In Section 20.3, however, we will only consider the case where the reflectivity

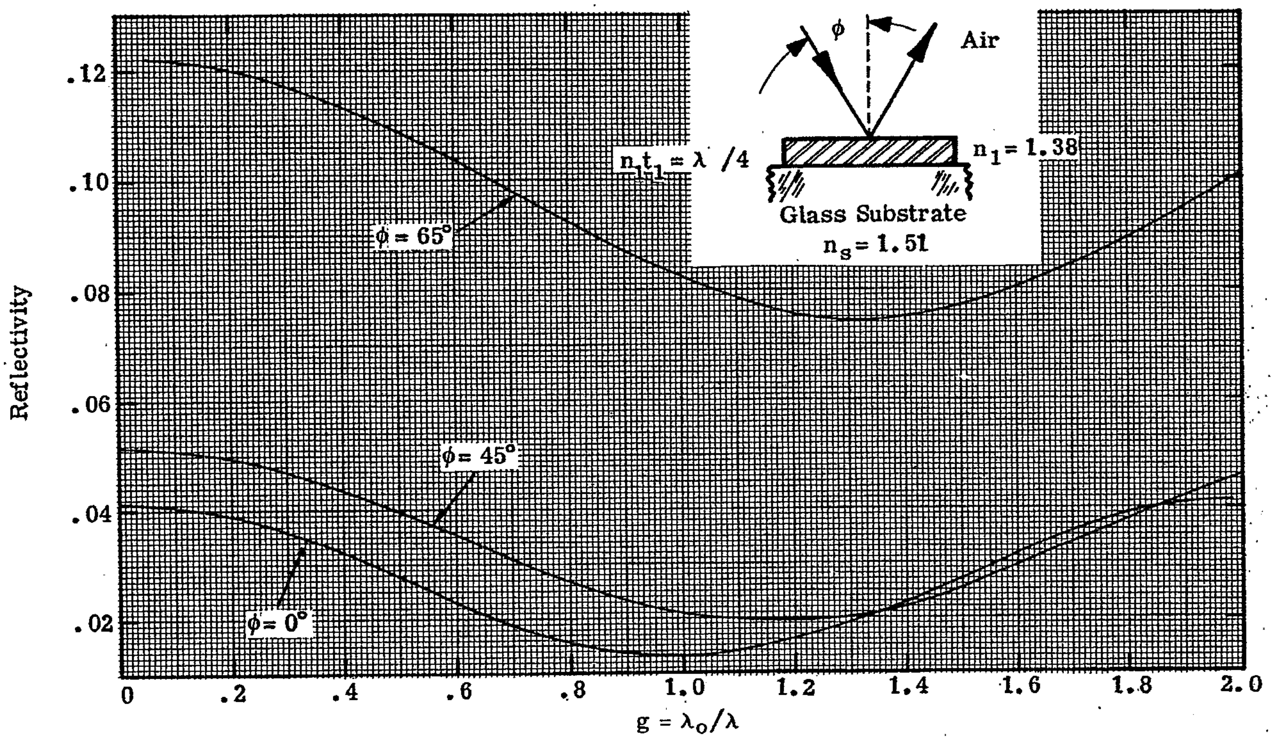


Figure 20.16 - Computed spectral reflectivity at various angles of incidence of a single-layer antireflection coating.

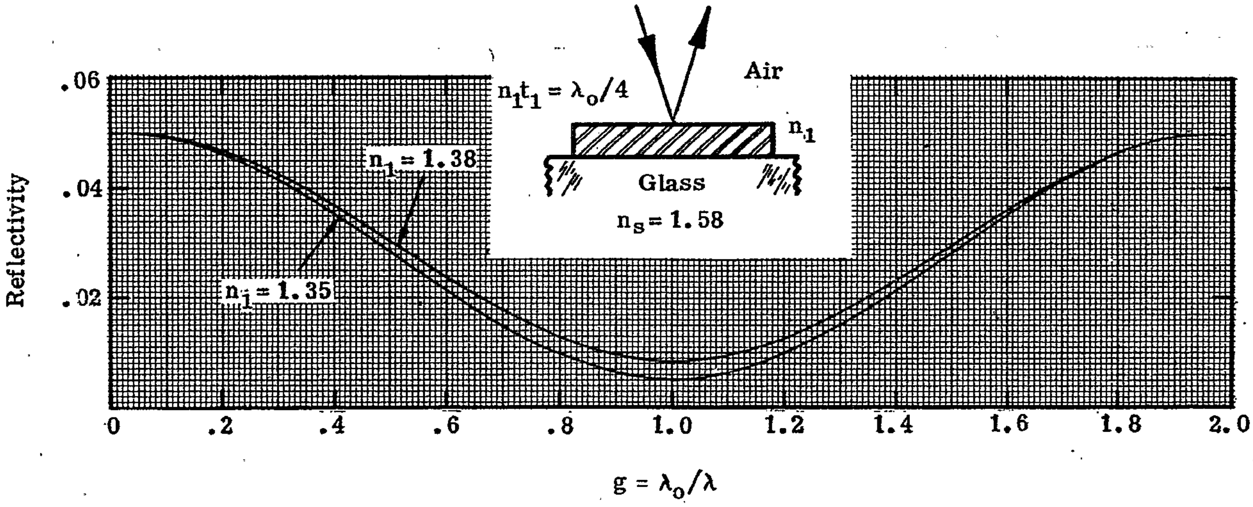


Figure 20.17 - Computed spectral reflectivity at normal incidence of a single-layer antireflection coating.

is a minimum, which occurs when

$$n_o > n_1 > n_s \quad \text{or} \quad n_o < n_1 < n_s . \qquad (25)$$

It is evident from Equation (24) that the reflectivity is zero when

$$n_1 = \sqrt{n_o \, n_s} \; . \qquad (26)$$

20.3.3.2 Coatings for substrates with low refractive index.

20.3.3.2.1 We will consider single-layer antireflection coatings for substrates which have a refractive index in the range from 1.5 to 1.7, which includes most of the optical materials which are commonly used in the visible, near-infrared and near-ultraviolet. Crown glass and fused quartz have a refractive index which is close to 1.50, while dense flint glass and sapphire have a refractive index of 1.70.

20.3.3.2.2 In order to obtain a coating which would produce zero reflectivity at a surface between air and glass of refractive index 1.51 at some particular wavelength, say 550 mμ, it is necessary to deposit a film which has an optical thickness of a quarter-wave, three-quarter wave, etc. at 550 mμ and a refractive index given in Equation (26), namely 1.23. No durable coating material has been found which has a refractive index of this low a value. Thus a single-layer coating on such a substrate is a compromise between a coating which is hard and durable and a coating which has an extremely low, if not zero, reflectivity. Two commonly used low-index films are magnesium fluoride (index 1.38) and cryolite (index 1.35). The cryolite film produces a lower reflectivity than the film of magnesium fluoride, because of the lower refractive index of the former. This advantage, however, is offset by the superior physical properties of the magnesium fluoride film. Cryolite films are soft and slightly water soluble, whereas magnesium fluoride films are quite hard and durable, especially when evaporated onto a hot substrate.

20.3.3.2.3 Figures 20.13, 20.14, and 20.15 show the computed spectral reflectivity of some antireflection coatings of both magnesium fluoride and cryolite on substrates of refractive index 1.51, 1.58, and 1.68. The reflectivity of a single uncoated surface at normal incidence in each of these cases is 4.12%, 5.05%, and 6.44%, respectively. The optical thickness of the film is $\lambda_o/4$ at $\lambda_o = 550$ mμ so that the minimum reflectivity is in the green region of the spectrum where the eye is most sensitive. Also, 550 mμ is in the center of the visible spectrum (i.e. 400 mμ to 700 mμ) on a wavelength scale. However, 550 mμ is not in the center of the visible spectrum on a wave number (frequency) scale and hence the reflectivity in Figures 20.13, 20.14, and 20.15 is higher at 400 mμ than at 700 mμ. The curves are not symmetrical, about λ_o because they are plotted on a wavelength rather than on a wave number (frequency) scale on the abscissa. The reflectivity of the cryolite coating is lower than the magnesium fluoride throughout the visible spectral region. The minimum reflectivity, R_m, decreases as the index of the substrate progresses from 1.51 to 1.68, because the condition in Equation (26) becomes closer to being satisfied.

20.3.3.2.4 The spectral reflectivity curves shown in Figures 20.13, 20.14, and 20.15 are useful because magnesium fluoride films are used so extensively to coat lenses in the visible spectral region. However, the reflectivity is shown only for a limited spectral region and for a film which has a quarter-wave optical thickness at a particular wavelength, namely 550 mμ. Suppose that a lens is designed to transmit radiant energy not only in the visible, but also in the near infrared to 950 mμ. If a designer wanted to know the reflectivity of a coated surface at 950 mμ, the data in Figures 20.13, 20.14, and 20.15 are of little use to him. Of course, he could compute the reflectivity at this wavelength from Equation (23), but it is much easier to obtain information from a graphical presentation. Suppose that a lens images radiant energy upon a detector which has a maximum response at a wavelength of 700 mμ. This means that the antireflection coatings should have a minimum reflectivity at 700 mμ and hence reflectivity versus wavelength plot depicted in Figures 20.13, 20.14, and 20.15 could not be directly used.

20.3.3.2.5 In order to circumvent the difficulties mentioned in the foregoing paragraph, it is useful to plot the reflectivity as a function of dimensionless parameter, $g = \lambda_o/\lambda$, which is proportional to frequency. Here λ_o is the wavelength at which optical thickness of the film is a quarter-wave. Figures 20.16, 20.17, and 20.18 depict such a plot, for a single layer of magnesium fluoride deposited on substrates of refractive index 1.51, 1.58, and 1.68, respectively. The R_{av} at non-normal incidence is shown for substrates of index 1.51 and 1.68. More extensive data for cryolite is not shown because it is not widely used as an antireflection coating because of the reasons mentioned in 20.3.3.2.2. When plotted versus "g", the reflectivity has even symmetry about $g = 1.0$ (at normal incidence).

20.3.3.2.6 Several examples are given on how these curves are used to compute the reflectivity at a wavelength λ of a single-layer coating which has $n_1 t_1 = \lambda_o/4$.

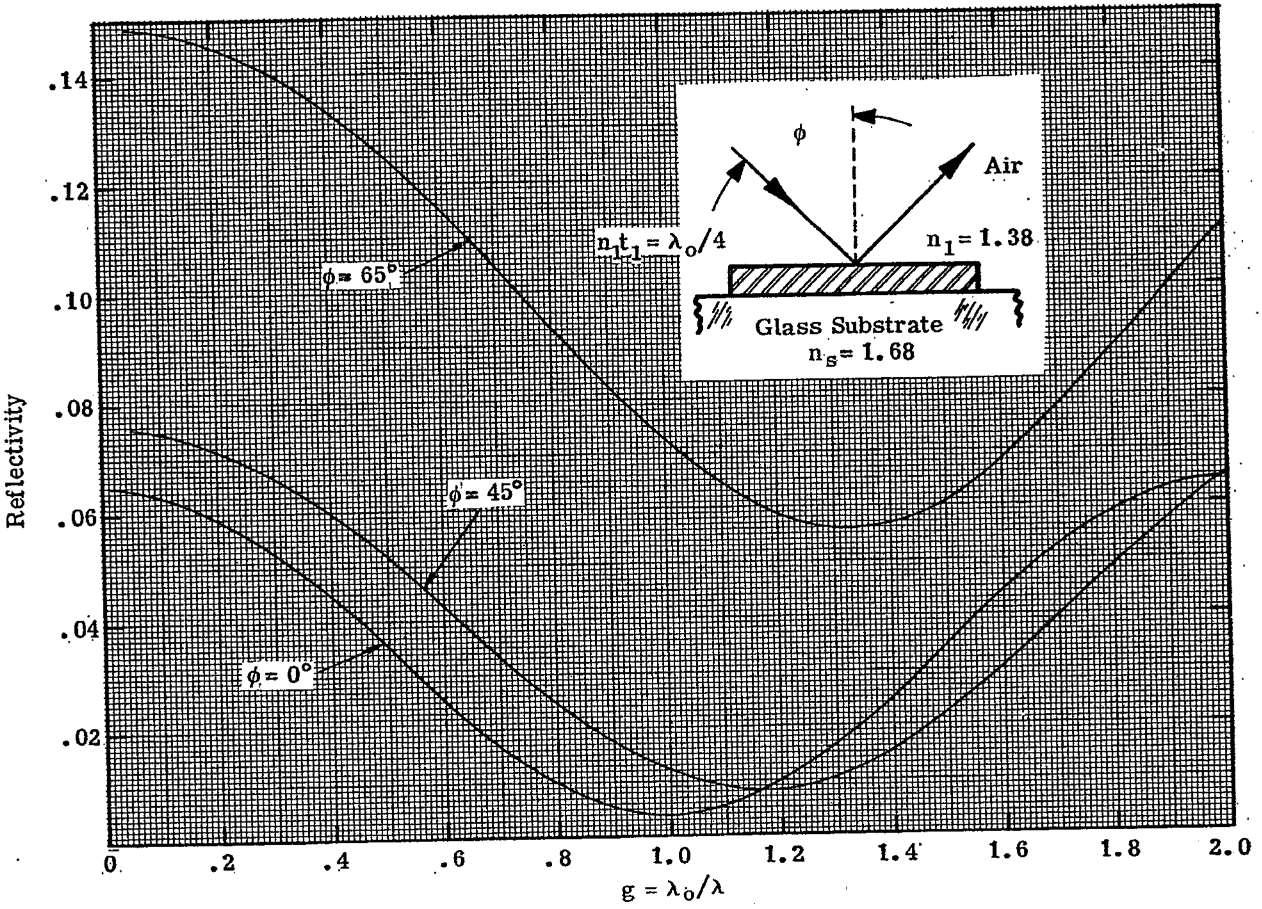


Figure 20.18 - Computed spectral reflectivity at various angles of incidence of a single-layer antireflection coating.

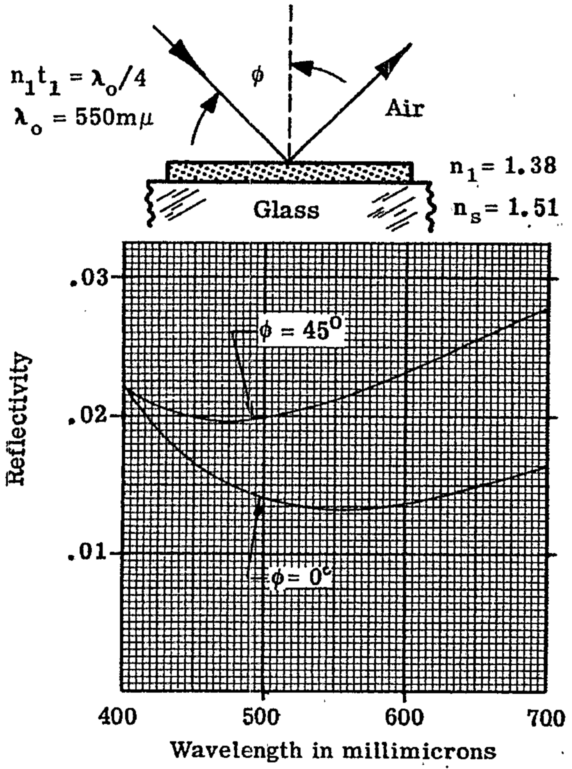


Figure 20.19 - Computed spectral reflectivity of a single layer at $\phi = 0^\circ$ and $\phi = 45^\circ$.

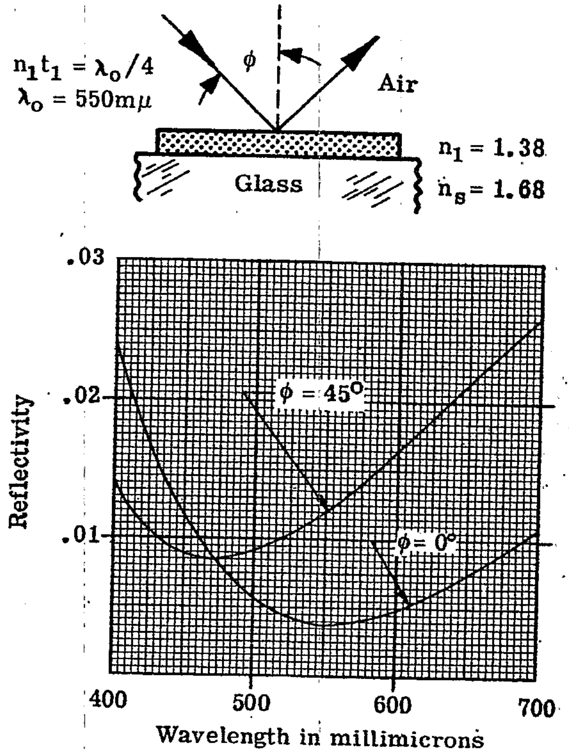


Figure 20.20 - Computed spectral reflectivity of a single layer at $\phi = 0$ and $\phi = 45^\circ$.

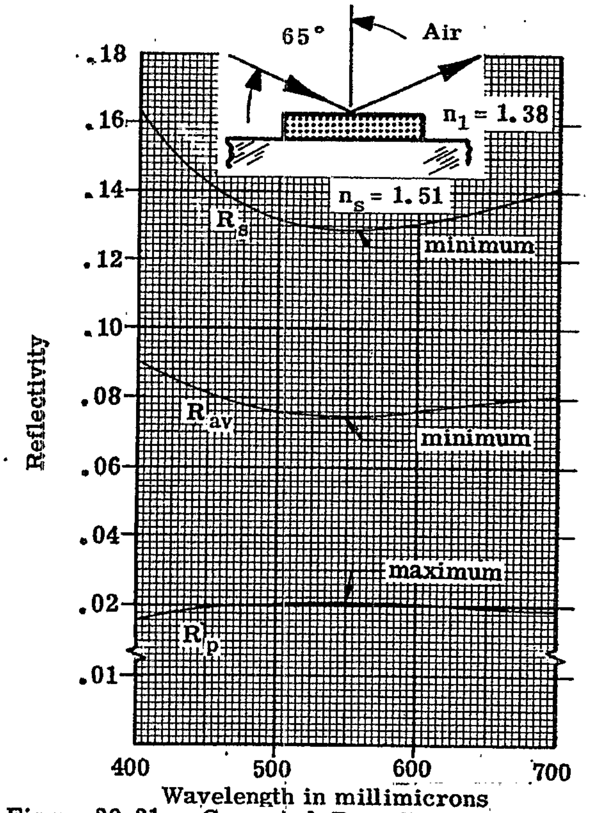


Figure 20.21 - Computed R_s, R_p, and $R_{av} = 1/2\ (R_p + R_s)$ at $\phi = 65^\circ$. The scale of the ordinate changes at 0.02.

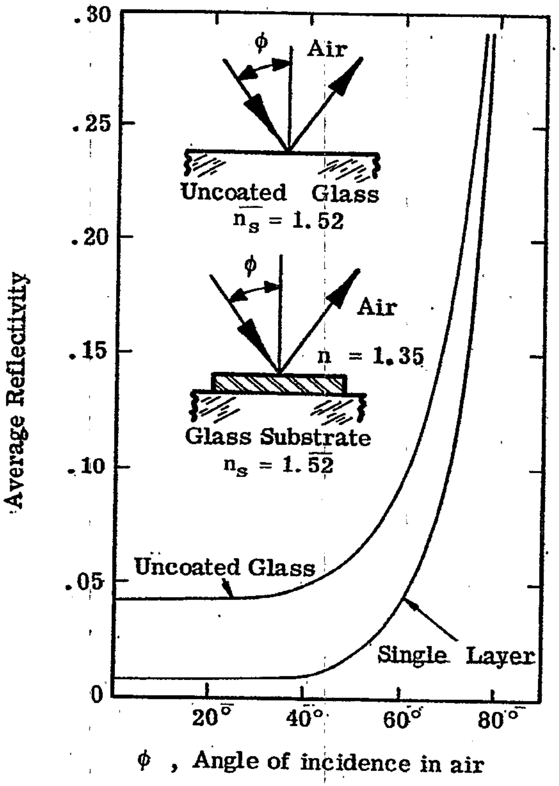


Figure 20.22 - Angle of incidence ϕ versus R_{av} of uncoated glass and the minimum R_{av} of the same glass coated with a single layer.

Example (a) A glass substrate of index 1.51 is coated with a single-layer magnesium fluoride antireflection coating with a minimum reflectivity at 700 mμ. What is the reflectivity at 950 mμ?

Solution: A quarter-wave film with $\lambda_o = 700$ mμ, has a minimum R at λ_o. Then at $\lambda = 950$ mμ, $g = \lambda_o/\lambda = 700/950 = 0.737$. From Figure 20.16, at $g = .737$ $R = .017$.

Example (b) For the same coating and substrate, what is the reflectivity at 400 mμ?

Solution: Here $g = \lambda_o/\lambda = 700/400 = 1.75$, and from Figure 20.16, $R = .038$.

20.3.3.2.7 The reflectivity versus "g" curves are periodic and at normal incidence the curve repeats at $g = 2.0$, 4.0, 6.0, etc. This fact can be used to find the reflectivity at values of "g" outside the range of the graphs. Example: For the thin film coating described in Paragraph 20.3.3.2.6, what is the reflectivity at 300 mμ? Here $g = \lambda_o/\lambda = 700/300 = 2.33$. Due to the periodicity, the reflectivity at $g = 2.33$ is the same as at $g = 0.33$ and from Figure 20.16 the reflectivity is 0.035.

20.3.3.3 Coatings for substrates with low refractive index, at non-normal incidence.

20.3.3.3.1 As is discussed in Section 20.1.6, Equations (23), (24), and (26) can be used at non-normal incidence, provided that an effective thickness is substituted for the optical thickness and the effective index appropriate to each plane of polarization is used. Regardless the angle of incidence, the reflectivity curve still has either a maxium or a minimum when the phase of retardation of the layer $\delta_1 = 90°$, 270°, etc. Because of the reasons cited in 20.1.6.4.2, the minimum in reflectivity curve shifts towards shorter wavelengths (i.e. the blue). This shift of the minimum reflectivity to shorter wavelengths is illustrated in Figures 20.19 and 20.20, which depict the spectral reflectivity of a layer of magnesium fluoride, ($n_1 t_1 = \lambda_o/4$, $\lambda_o = 550$ mμ) on glass substrates of index 1.51 and 1.68 respectively. At non-normal incidence, R_{av} is shown. The minimum reflectivity shifts from 550 mμ at $\phi = 0$ to $\lambda_2 = 465$ mμ at $\phi = 45°$. This shift can also be determined from the graph of the effective thickness versus ϕ in Figure 20.8. From this graph we find that $(n\ t)_{effective} = n_1 t_1 \cos\theta_1 = 0.859\, n_1 t_1$ for a film index of 1.38. Given λ_o of 550 mμ, we determine that $\lambda_2 = (.859)(550) = .472$ mμ.

20.3.3.3.2 The effective indices at $\phi = 65°$ of the substrate, film, and incident medium of a single-layer MgF_2 coating on glass, are shown in Table 20.3. Its spectral reflectivity in each plane of polarization and the R_{av} is shown in Figure 20.21. The optical thickness of the film is made thicker than $\lambda_o/4$ at normal incidence so that at $\phi = 65°$ the minimum reflectivity occurs at $\lambda_o = 550$ mμ. It is interesting to note that R_p attains a maximum rather than a minimum at λ_o. The reason for this is seen in Table 20.3. The effective indices in the "p" plane of polarization do not satisfy the condition for a minimum stated in Equation (25). However, the R_s curve drops to a sharp minimum at λ_o and hence the R_{av} has a minimum, rather than a maximum, at λ_o. Figures 20.16 and 20.18 show the average reflectivity at $\phi = 45°$ and $\phi = 65°$ of a film of refractive index 1.38 deposited on glass of index 1.51 and 1.68, respectively. The optical thickness is $\lambda_o/4$ at $\phi = 0$. These curves have even symmetry about their minimum. For example, the curve in Figure 20.16 at $\phi = 65°$ has a minimum at $g = 1.32$ and hence the reflectivity is the same at $g = 2.12$ and $g = 0.52$.

20.3.3.3.3 Graphs of the average reflectivity, such as are shown in Figures 20.16, 20.18, 20.19 and 20.20, are useful in the case where the incident light is unpolarized and where the detector is not sensitive to polarization, as for example a photographic plate. If the incident light is polarized, or if polarization has been introduced by other elements of an optical system, such as prisms and beam splitters, then it is necessary to compute the reflectivity in each plane of polarization separately, as in Figure 20.27.

20.3.3.3.4 Single-layer antireflection coatings (which satisfy Equation (25)) on glass always decrease the average reflectivity to lower values than the uncoated surface. This is illustrated in Figure 20.22 which shows a plot of the angle of incidence versus R_{av} of an uncoated glass surface, and the minimum reflectivity of the same substrate covered with a single layer of refractive index 1.35. At any angle of incidence, whether it be 20° or 80°, the coated surface has a lower R_{av} than the bare substrate. This subject of antireflection coatings at non-normal incidence is treated in more detail in references 25 and 26.

20.3.3.4 Coatings for substrates with a higher refractive index.

20.3.3.4.1 Single-layer antireflection coatings for substrates with a higher refractive index are considered, which includes materials which are used principally in the infrared, such as arsenic trisulfide glass (index 2.40), silicon (index 3.48), and germanium (index 4.045). The refractive index of all of these materials changes with wavelength. The foregoing values are representative of some mean values in the infrared and even though the calculations do not account for the dispersion in the refractive index, they give some idea of what can be accomplished in the way of antireflection coatings for these materials.

LAYER	INDEX AT NORMAL INCIDENCE	EFFECTIVE INDEX AT $\phi = 65^\circ$	
		POLARIZATION P	POLARIZATION S
Incident medium (Air)	1.00	2.366	0.423
Magnesium fluoride	1.38	1.83	1.041
Glass substrate	1.51	1.89	1.208

Table 20.3 - The effective indices at $\phi = 65^\circ$ incidence of a single-layer antireflection coating.

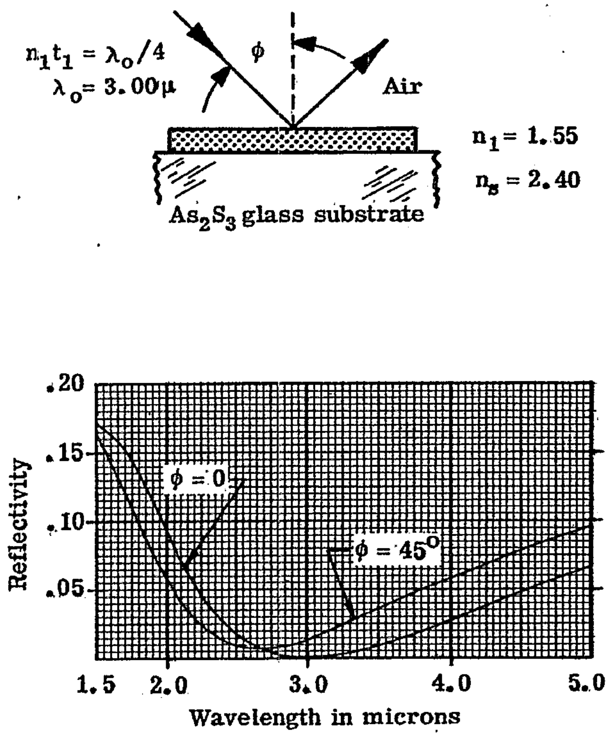


Figure 20.23 - Computed spectral reflectivity of a single-layer antireflection coating at $\phi = 0$ and $\phi = 45^\circ$.

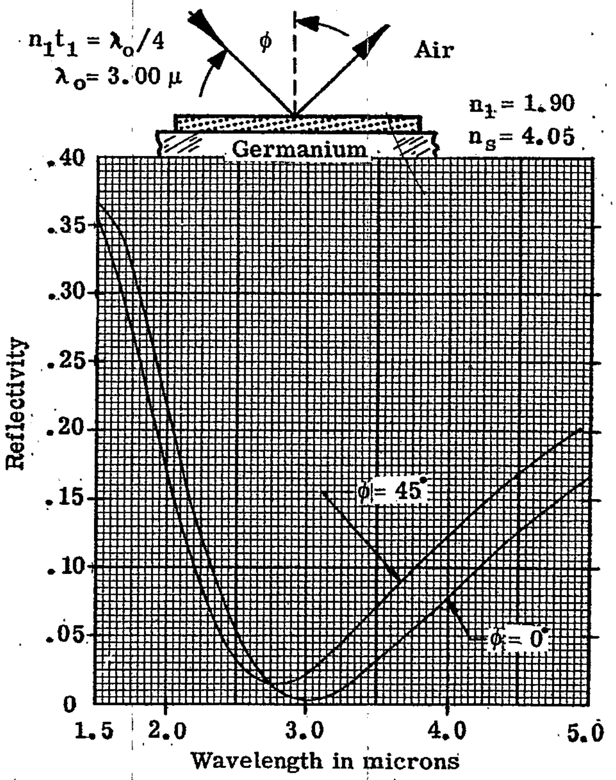


Figure 20.24 - Computed spectral reflectivity of a single-layer antireflection coating at $\phi = 0$ and $\phi = 45^\circ$.

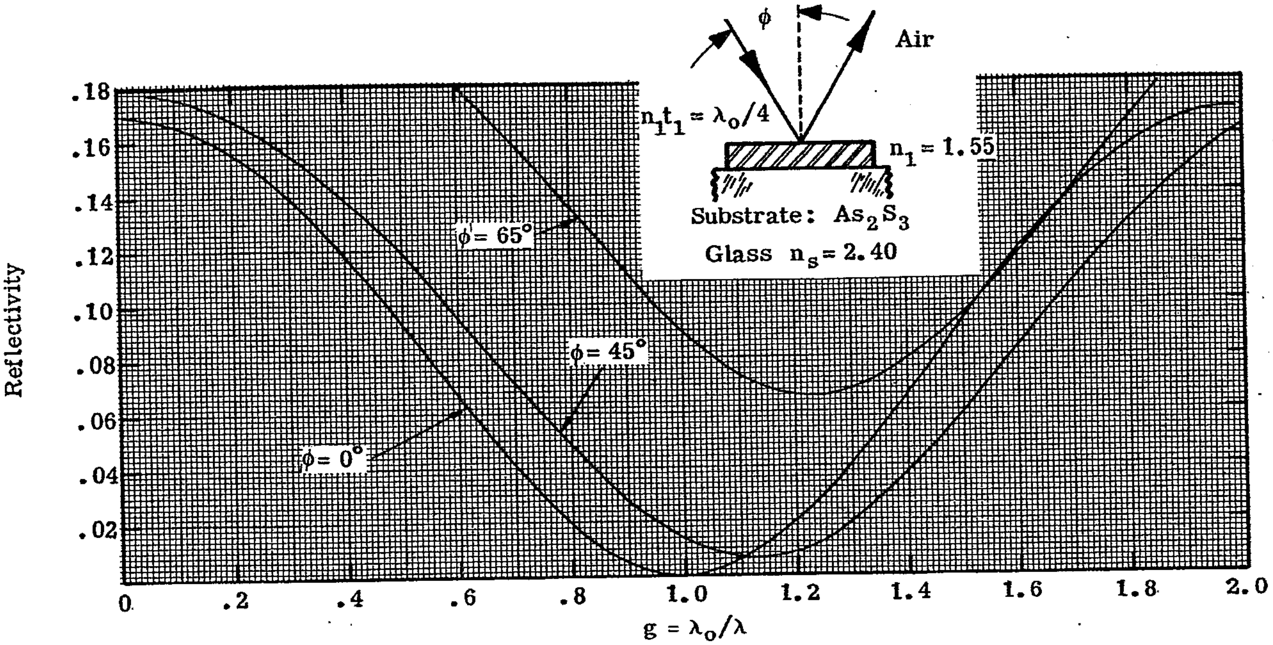


Figure 20.25 – Computed spectral reflectivity of a single-layer antireflection coating at ϕ = 0, 45° and 65° .

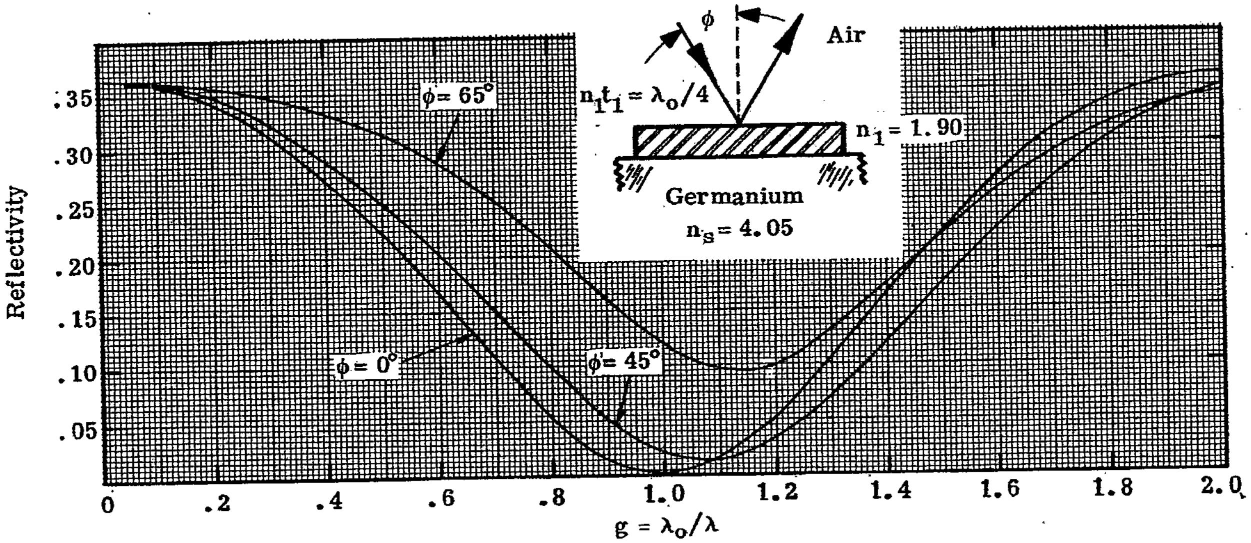


Figure 20.26 – Computed spectral reflectivity of a single-layer antireflection coating at ϕ = 0, 45° and 65° .

20.3.3.4.2 The difference between the coatings for low-index substrates and these higher index materials is that in the latter case thin film materials are available which satisfy Equation (26). Thus the antireflection layer on arsenic trisulfide glass should have a refractive index of $(2.40)^{1/2} = 1.55$, and so on for the other substrates. Figures 20.23 and 20.24 show single layer antireflecting coatings with a quarter-wave optical thickness of 3.0 μ, on substrates of arsenic trisulfide glass and germanium. The reflectivity of the uncoated surface is 0.17 and 0.364 respectively. It should be noted that in practice the reflectivity curve in Figure 20.24 is higher than shown in the region from 1.5 to 1.8 μ, because the absorption constant of the germanium is increasing. Figures 20.25 and 20.26 show a plot of the reflectivity versus "g". In each case the optical thickness of the films is $\lambda_o/4$. These curves can be used to find the reflectivity at various wavelengths of a film which has a quarter-wave optical thickness, say at 4.0 μ, in the manner described in Section 20.3.3.2.6.

20.3.3.5 Coatings for substrates with a higher refractive index, at non-normal incidence.

20.3.3.5.1 The behavior at non-normal incidence of coatings for substrates with a low refractive index was discussed in 20.3.3.3. That discussion applies equally well to these coatings for high-index substrates. The main difference is that the angle shift of the reflection minima to shorter wavelengths is considerably less for the high-index coatings. This is merely a manifestation of the fact that the change in the effective thickness is much less for high-index films than for the low index films, as one can see from Figure 20.8.

20.3.3.5.2 Figure 20.27 shows the spectral reflectivity at 65° in the two planes of polarization and also the R_{av} of a single layer coating on germanium. As in the case of the film shown in Figure 20.21, the effective indices of the incident medium, film, and substrate in the "p" plane of polarization do not satisfy Equation (25) and hence the reflectivity attains a maximum rather than a minimum. However, in the case of the germanium film, R_s is less than R_p. This means that at $\phi = 65°$ the polarizing effect of this coated plate is exactly opposite to that produced by an uncoated dielectric surface, which satisfies the condition that $R_s > R_p$, for all values of $\phi > 0$. This coating on germanium could be used to compensate for the polarization introduced by other uncoated surfaces in an optical system. Figures 20.25 and 20.26 show the average reflectivity of these coatings at various angles of incidence. Hass[27, 28] and his co-workers have measured the transmittance in the infrared of some antireflected high-index substrates.

20.3.3.6 Coatings with a higher order of interference. As was pointed out in 20.3.3.1.2, a minimum in the reflectivity of a single-layer coating occurs when the optical thickness of the coating is $\lambda/4$, $3\lambda/4$, $5\lambda/4$, $m\lambda/4$, where the order number "m" is an odd integer. Hitherto we have only shown coatings which show a first-order interference minimum, that is, for $m = 1$. A minimum reflectivity will also occur for higher order interference coatings, such as films which have an optical thickness of three-quarter or five-quarter waves. Figure 20.28 shows the spectral reflectivity of these thicker films deposited on a glass substrate. As one might expect, the reflectivity rises sharply on either side of the minimum at 550 mμ. Hence, there is little advantage to using such higher order interference coatings, with the following exception. In an infrared optical system which is designed to transmit a narrow band of wavelength, some additional attenuation of wavelengths outside of the desired range could be obtained by using higher order interference antireflection coatings.

20.3.3.7 Analogy with electrical transmission lines. In this section we have considered the problem of light impinging upon a substrate of index n_s from an incident medium of index n_o. The problem has been to select the film of proper optical thickness and refractive index so that the reflectivity is reduced to zero. The analogous problem in transmission line theory to match a load of admittance n_s to a transmission line of characteristic admittance n_o so that there will be no standing waves. It is shown in many texts on transmission line and microwave theory that a "quarter-wave transformer" or "quarter-wave matching line" is required to do this. The electrical length of the line should be a quarter-wave and the admittance, n_1, of the line should satisfy Equation (26).

20.3.4 Double-layer coatings.

20.3.4.1 Types of coatings. It is often desirable to use double-layer antireflection coatings because in certain cases these coatings have a lower reflectivity over a wider spectral region than do single-layer coatings. The following types of coatings will be considered:

(1) Double-quarter, single minimum.

(2) Double-quarter, double minimum.

(3) Quarter-half.

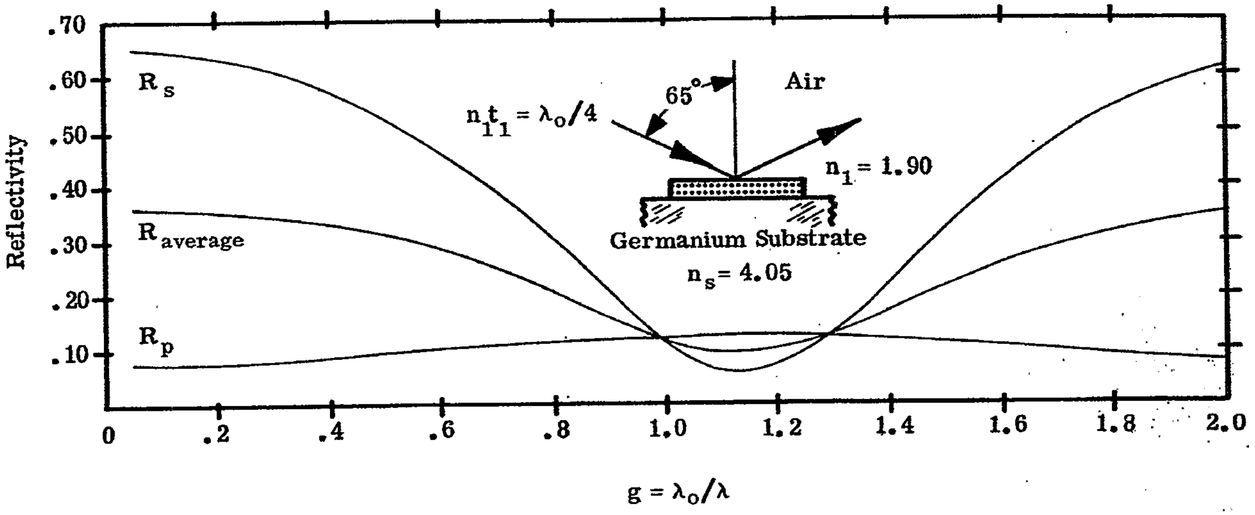


Figure 20.27 Computed R_s, R_p and $R_{av} = 1/2\ (R_s + R_p)$ at $\phi = 65°$.

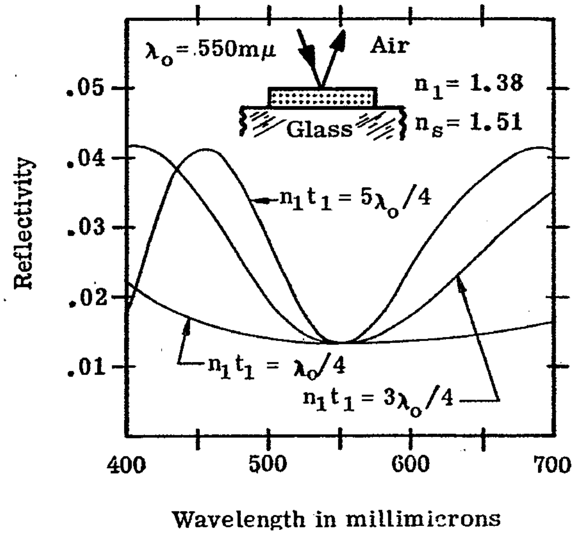


Figure 20.28 Computed spectral reflectivity of single-layer antireflection coatings of various thickness.

Type of Coating	n_1	n_1t_1	n_2	n_2t_2
Single Quarter	1.38	$\lambda_0/4$	—	—
Double Quarter	1.38	$\lambda_0/4$	1.70	$\lambda_0/4$
Quarter-Half	1.38	$\lambda_0/4$	1.80	$\lambda_0/2$

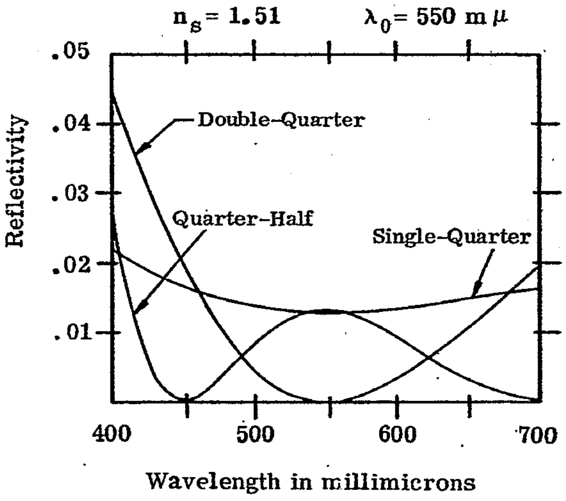


Figure 20.29 Computed spectral reflectivity of single and double-layer antireflection coatings on a glass substrate.

20.3.4.2 Basic equations.

20.3.4.2.1 It is possible to write an equation for the reflectivity of a double-layer coating, but there is little advantage because the equation is lengthy and cumbersome.[29] For our purposes, it is preferable to use the vector addition of amplitude or the characteristic matrix method to compute the reflectivity.

20.3.4.2.2 In the special case where the optical thickness of each film is a quarter-wavelength, that is $n_1 t_1 = n_2 t_2 = \lambda/4$, then $\delta_1 = \delta_2 = 90°$ and the characteristic matrices (Equation (6)) have non-zero elements only along the antidiagonal. The matrix product (Equations (7), (8)) becomes:

$$\begin{bmatrix} A & jB \\ jC & D \end{bmatrix} = \begin{bmatrix} 0 & jn_1^{-1} \\ jn_1 & 0 \end{bmatrix} \begin{bmatrix} 0 & jn_2^{-1} \\ jn_2 & 0 \end{bmatrix} . \qquad (29)$$

After taking the matrix product and substituting into Equations (9) and (10), we obtain for the minimum (or maximum) reflectivity

$$R = 1 - T = \left[\frac{n_o n_2^2 - n_s n_1^2}{n_o n_2^2 + n_s n_1^2}\right]^2 . \qquad (30)$$

Thus for R to be zero, the condition must be satisfied that:

$$\left(\frac{n_2}{n_1}\right) = \left(\frac{n_s}{n_o}\right)^{1/2} . \qquad (31)$$

The condition for zero reflectivity of a double quarter, single minimum type of coating only involves the ratio of the indices of the two films. For example, suppose glass of index 1.51 is coated with a double-quarter-wave coating with the following indices:

$$n_1 = 1.38 \qquad n_2 = 1.69 \qquad (32a)$$

$$n_1 = 1.65 \qquad n_2 = 2.03 \qquad (32b)$$

In both cases Equation (31) is satisfied and the reflectivity is zero at $\lambda = \lambda_o$. However, it can be shown that the reflectivity rises quite sharply on either side of the minimum in the case of Equation (32b), whereas the spectral range over which the reflectivity has a low value is much larger in the case of Equation (32a), and thus Equation (32a) is the better coating to use. Hence it is preferable to use as low index materials as possible.

If the indices satisfy Equation (31), then this is called a double-quarter, single minimum coating. The single minimum means that it has only one reflectivity minimum for a given order of interference. For example, the plot of the reflectivity versus λ_o/λ of the double-quarter coating (curve I in Figure 20.30) has one minimum for the first order interference for $0 \leq g \leq 2.0$, another minimum in the second order for $2.0 \leq g \leq 4.0$, and so on.

20.3.4.2.3 Another type of coating which contains quarter-wave layers has a maximum R at λ_o and two minimum symmetrically spaced about λ_o on a frequency scale. Such a coating is called a double-quarter double minimum coating. A reflectivity curve for a typical coating is shown in Figure 20.35. There are several methods of determining the relationship between the indices of the films of such a coating, such as the solution of complicated algebretic equations[27, 29, 30] or alternatively the use of vector diagrams, as done in Figure 20.5. One type of double-quarter double minimum coating is obtained when the indices satisfy the equations

$$n_1^3 = n_o^2 n_s , \qquad n_2^3 = n_o n_s^2 . \qquad (33)$$

The foregoing equations can be reduced to

$$n_1 n_2 = n_o n_s . \qquad (34)$$

Equations (33) and (34) are derived by Berning[30].

20.3.4.2.4 The spectral reflectivity curve of the quarter-half coating is similar to the curve of the double-quarter, double minimum coating mentioned in the foregoing paragraph, to the extent that it has a maximum at λ_o and two minimum spaced equally about λ_o on a frequency scale. The spectral reflectivity curve of such a coating is depicted in Figure 20.30 (curve III). It is required that the half-wave layer must have a

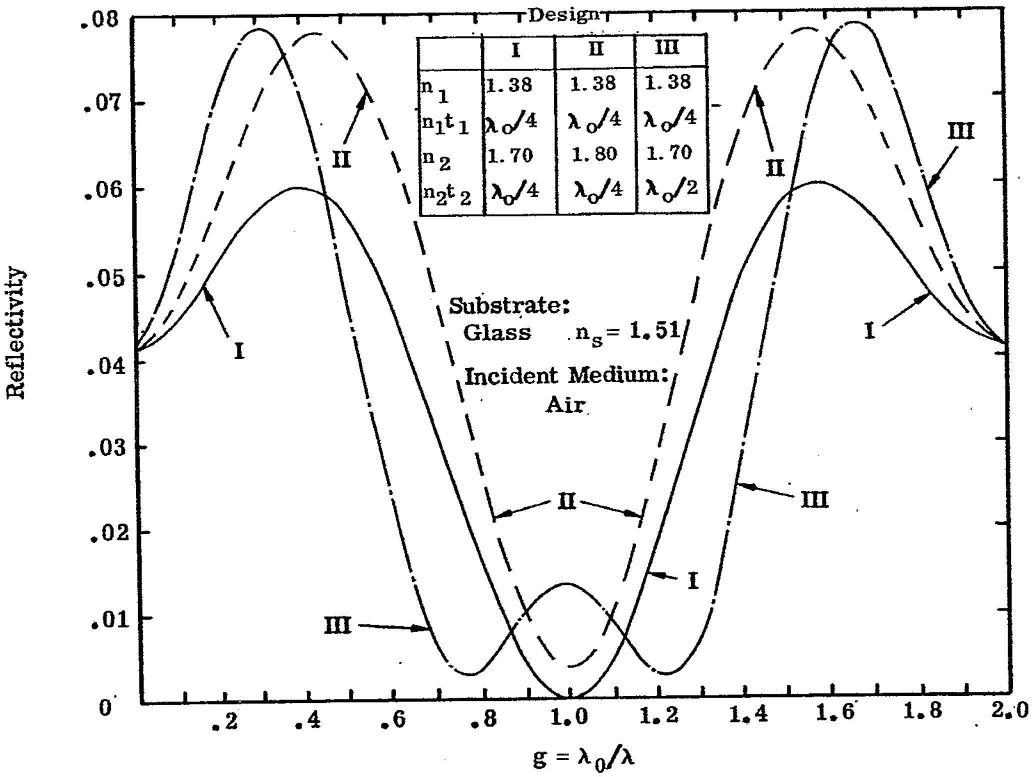


Figure 20.30 – Computed spectral reflectivity of single and double-layer antireflection coatings on a glass substrate.

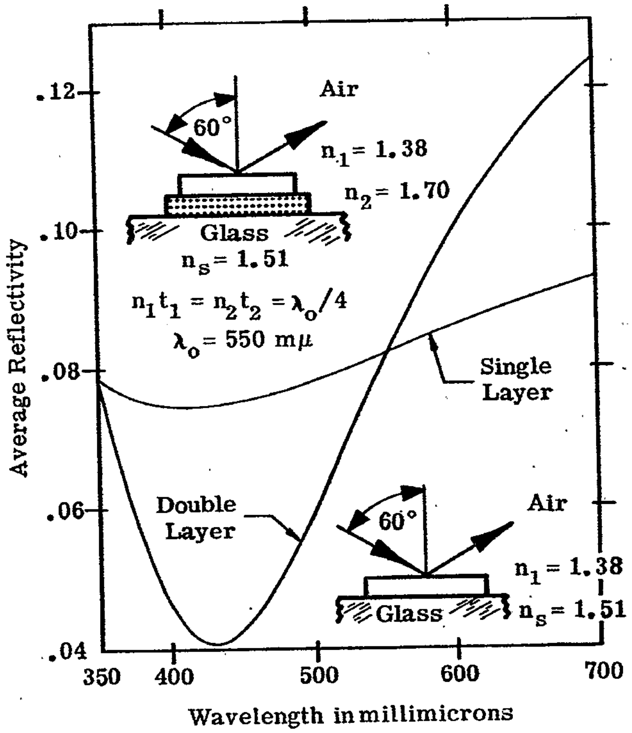


Figure 20.31 - Computed average spectral reflectivity of a single-layer and a double-layer coating at $\phi = 60^\circ$.

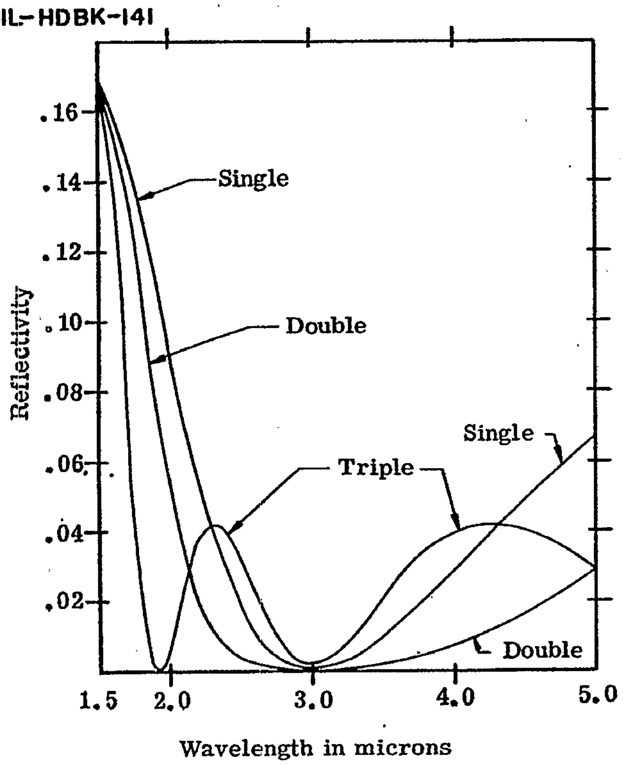


Figure 20.32 – Computed spectral reflectivity of the antireflection coatings whose designs are depicted in Fig. 20.33.

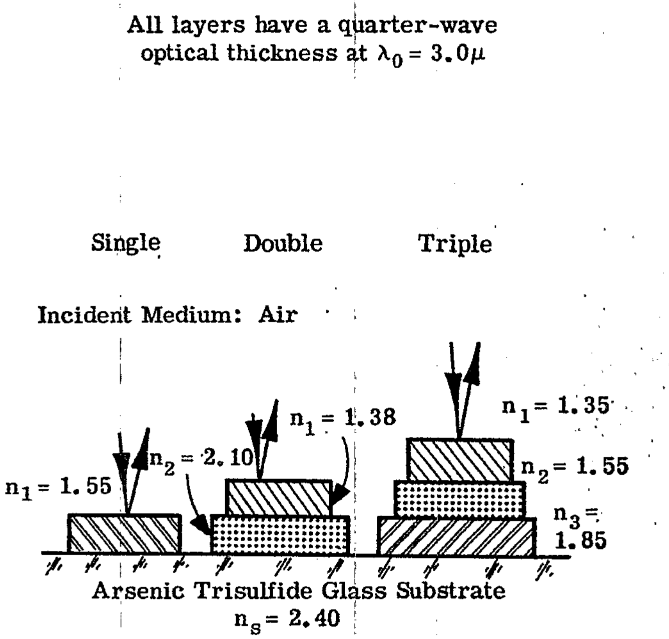


Figure 20.33 — The designs of single, double, and triple-layer antireflection coatings whose reflectivity curves are shown in Figure 20.32.

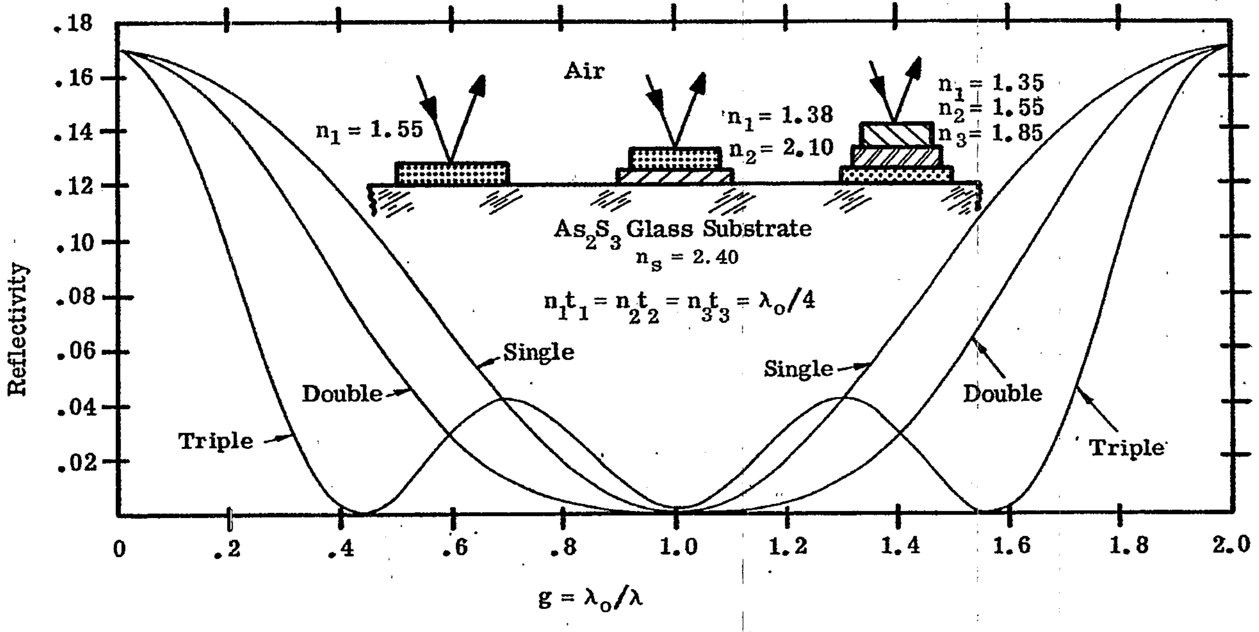


Figure 20.34 – Computed spectral reflectivity of antireflection coatings at normal incidence.

higher refractive index than the substrate. The refractive indices of the two layers can be determined either from a vector diagram (i.e. Figure 20.5) by adjusting the lengths of the vectors so that the vector polygon should close at some value of δ_1 and δ_2. The length of the vectors is proportional to the Fresnel amplitude coefficients and hence one can solve for the indices of the layers. The refractive indices can also be found by solving transendental equation.[29] Whatever indices are chosen, the reflectivity when $\lambda = \lambda_o$ is determined easily from Equation (24) because the half-wave layer is absentee.

20.3.4.3 Double-layer antireflection coatings for low index substrates.

20.3.4.3.1 In this subsection, the term low index substrate is used in the same sense as in 20.3.3.2. The fact that the lowest index film material which is available has an index of 1.35 means that double quarter, double minimum coatings cannot be produced for substrates in this index range.

20.3.4.3.2 The double quarter curve in Figure 20.29 is the spectral reflectivity of a double-quarter, single minimum coating on glass of refractive index 1.51. The optical thickness of the films is a quarter-wave at 550 mμ and because the indices satisfy Equation (31), the reflectivity at this wavelength is zero. Such a coating could be manufactured using magnesium fluoride as the low index film and silicon monoxide as the high index film. Unfortunately, when the index of the silicon monoxide is this large (see Note 4 in Table 20.2), it has a slight absorption in the blue and hence the films are yellowish in appearance. On the same graph is shown for comparison, the reflectivity of a single quarter-wave coating of magnesium fluoride. The reflectivity of the double-quarter coating is below the single layer in the green, but rises considerably above it at 400 mμ. Figure 20.30 shows the reflectivity of the double-quarter coating on a frequency scale. At certain wavelengths, the reflectivity is greater than that of uncoated substrates. Shown also in Figure 20.30 is the reflectivity of a double-quarter coating composed of films of indices 1.38 and 1.80. This shows what happens to the reflectivity when the relationship in Equation (31) are not precisely satisfied. Hass[29] shows many computed reflectivity curves in which variations have been made in both the thickness and refractive indices from the optimum condition specified in Equation (31).

20.3.4.3.3 Figure 20.29 shows the spectral reflectivity of a quarter-half coating. As was stated in 20.3.4.2.4, at λ_o (550 mμ) the half-wave layer is absentee and the reflectivity is the same as the single quarter-wave layer of index 1.38. Since the two minimum are equally spaced about the maximum on a frequency scale, they are unequally spaced on the wavelength scale. Figure 20.30 shows the spectral reflectivity of the same configuration, but on a λ_o/λ scale.

20.3.4.3.4 The limitation on the film index mentioned 20.3.4.3.1 means that the reflectivity of both the double-quarter and the quarter-half antireflection coatings exceeds the reflectivity of the uncoated substrate at some wavelengths, as is shown in Figure 20.30. This is not true of the single-layer low index coating, whose reflectivity never exceeds that of the uncoated substrate. The question as to which type of coating to use depends upon the range of wavelengths over which the reflectivity is to have a low value. If the range is narrow, a double-quarter coating might be preferable. However, if the range is quite large, then even though the single-layer coating does not achieve as low a reflectivity as the double-quarter coating in certain spectral regions, the better overall performance of the single layer over a wide range of wavelengths would make it preferable.

20.3.4.4 Double-layer antireflection coatings for low index substrates at non-normal incidence. At non-normal incidence the reflectivity of coatings which have two or more layers is influenced by the fact that if the optical thicknesses of the layers are matched at normal incidence, they are no longer matched at any other angle. This point is discussed in more detail in 20.1.6.3 and 20.1.6.4. If the layers are matched at some angle of incidence, ϕ, then one can compute the reflectivity at $\lambda = \lambda_o$ by substituting the effective index appropriate to each plane of polarization into Equation (30). In the case of double layer coatings, there are many possible combinations of incident angles and matched or mismatched layers which can be considered and hence a complete analysis of the behavior of the double-layer coatings of non-normal incidence would be quite lengthy. As an illustration of a typical case, Figure 20.31 shows the spectral reflectivity of a double-quarter coating on glass at 60° incidence. Both layers have a quarter-wave optical thickness at normal incidence at 550 mμ. The reflectivity minimum has shifted to the blue and notwithstanding the fact that the optical thickness of the layers is mismatched, the minimum reflectivity is still considerably lower than the minimum of the single-layer coating.

20.3.4.5 Double-layer antireflection coatings for high index substrates.

20.3.4.5.1 In this subsection, the term high index substrate is used in the same sense as in 20.3.3.4.1. For high-index substrates, coating materials are available which not only satisfy Equations (31), (33), or (34), but also the relationship:

$$n_o < n_1 < n_2 < n_s . \tag{35}$$

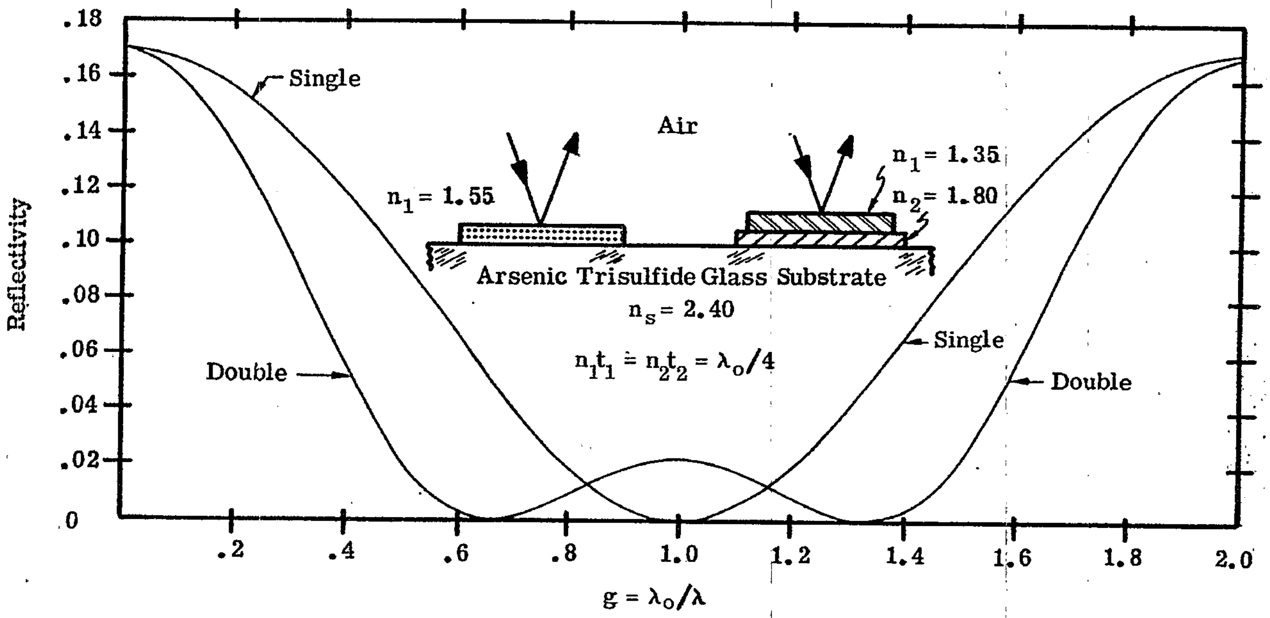


Figure 20.35 – Computed spectral reflectivity of antireflection coatings at normal incidence.

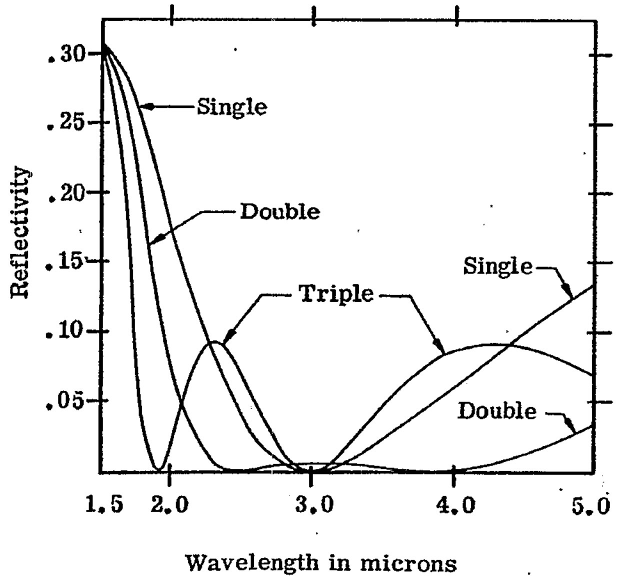


Figure 20.36 – Computed spectral reflectivity of the antireflection coatings whose designs are depicted in Fig. 20.37.

All layers have a quarter-wave optical thickness at $\lambda_0 = 3.0\mu$

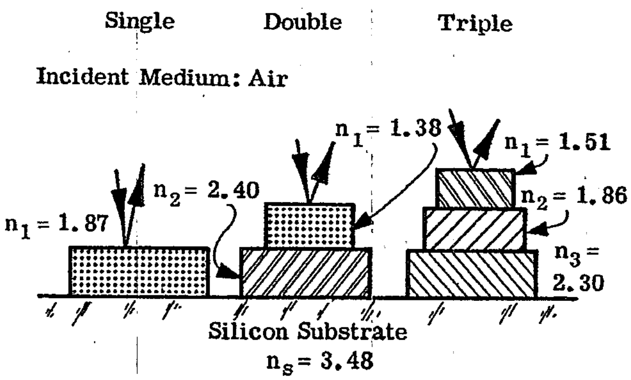


Figure 20.37 – The designs of single, double, and triple-layer antireflection coatings whose reflectivity curves are shown in Figure 20.36.

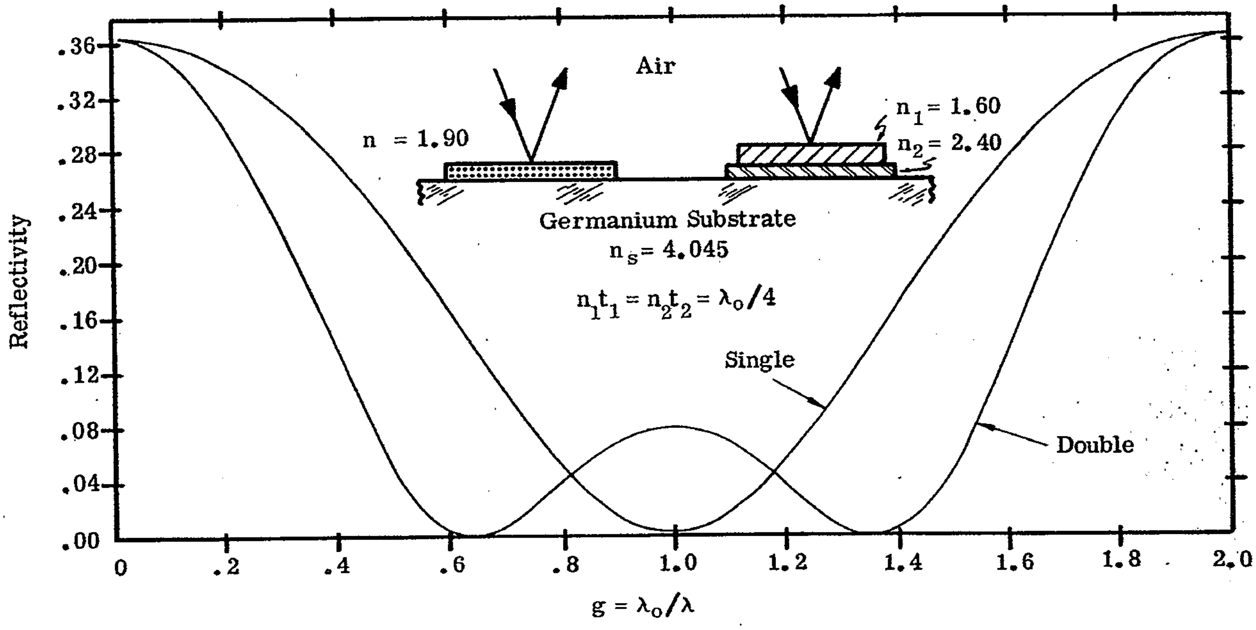


Figure 20.38 – Computed spectral reflectivity of antireflection coatings at normal incidence.

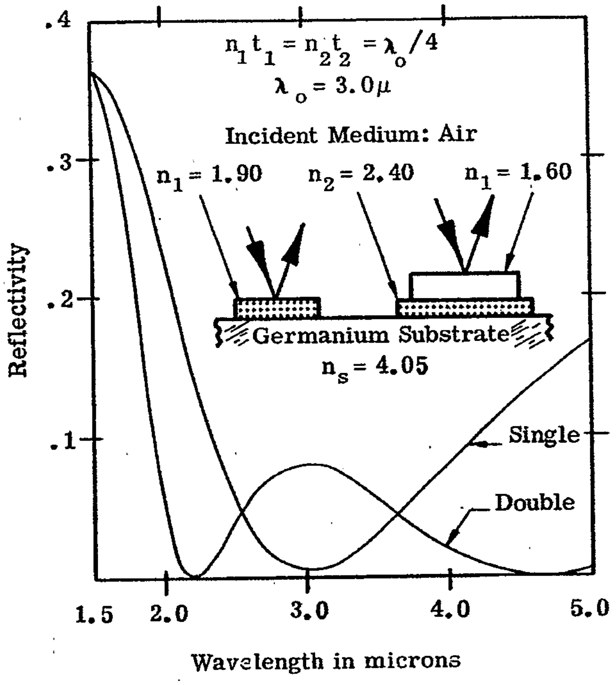


Figure 20.39 – Computed spectral reflectivity of antireflection coatings at normal incidence.

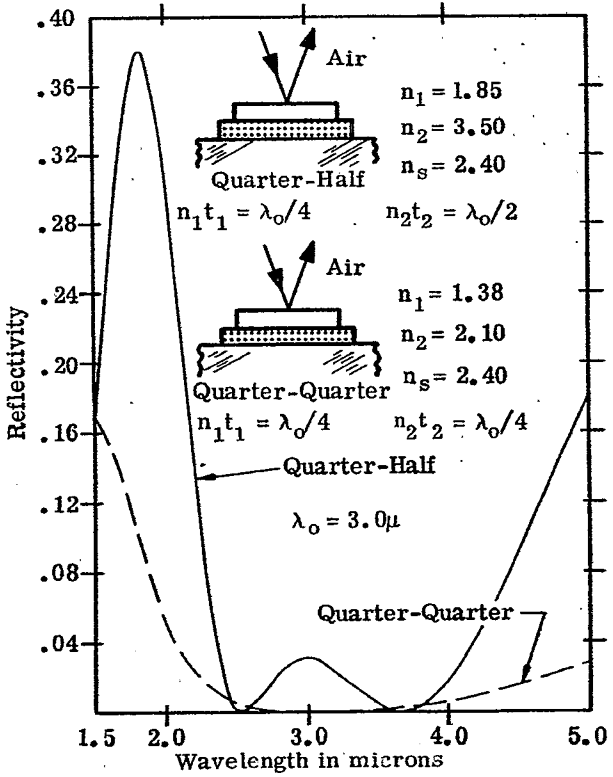


Figure 20.40 – Computed spectral reflectivity of antireflection coatings at normal incidence.

In this case a wide region of low reflectivity is achieved and the reflectivity does not exceed the reflectivity of the uncoated substrate. This fact is illustrated in Figure 20.34, which depicts a double-quarter, single minimum coating on a substrate of arsenic trisulfide glass.

20.3.4.5.2 With higher index substrates it is possible to use film materials which satisfy Equation (33) and hence it is possible to produce double-quarter, double minimum coatings. Figure 20.35 shows the spectral reflectivity of such a coating on arsenic trisulfide glass; the indices of the layers satisfy Equations (33) and (34). Figure 20.36 shows a double-quarter, double minimum coating on a silicon substrate in which the indices of the films satisfy Equation (34) to a fair approximation. Figures 20.38 and 20.39 show the spectral reflectivity of a double-quarter, double minimum coating on a germanium substrate. The indices of the films in this coating satisfy Equations (33) and (35).

20.3.4.5.3 Figure 20.40 shows the spectral reflectivity of a half-quarter coating on a substrate of arsenic trisulfide glass. The layer with the half-wave optical thickness is silicon, which has a higher refractive index than the substrate. The reflectivity rises considerably above that of the uncoated substrate and hence there seems to be little advantage of this type of coating over the double-quarter coatings.

20.3.4.5.4 Figure 20.41 shows the spectral transmittance of a germanium plate with both sides antireflected by a two-layer coating.[28] With the double-quarter single minimum coating, the ratio of the index of the silicon film to the index of the didymium fluoride film closely approximates the ratio specified in Equation (31). Figure 20.42 shows the transmittance of a silicon plate which has been coated in a similar manner.[28] The index of the cerium dioxide film and the magnesium fluoride film of this double-quarter double minimum coating satisfy Equation (34). The double reflectivity minimum of this type of coating are manifested in the transmission maxima at 1.8 and 2.7 μ. The bump in the curve near 3 μ is due to the water adsorbed on the coating. In both cases, the spectral region in which the transmission exceeds 0.9 is wider for these double-layer coatings than for a single layer coating.

20.3.5 Triple-layer antireflection coatings.

20.3.5.1 Types of coatings. The number of types of three-layer antireflection coatings which can be produced is legion, because there are six parameters which can be varied - three thicknesses and the three refractive indices. The discussion is confined to two types of coatings:

(1) The quarter-half-quarter.

(2) The triple-quarter, triple minimum.

20.3.5.2 Basic equations. Triple-layer coatings are usually designed by the use of vector diagrams (Section 20.1.5) or with the more sophisticated methods of Section 21.7.4. A few of the many conditions for zero reflectivity of three-layer coatings are derived by Berning.[30] One of the conditions for a stack of three quarter waves to attain a triple reflectivity minimum is:

$$n_1 n_3 = n_o n_s \quad (36a)$$

$$n_2^2 = n_o n_s \quad . \quad (36b)$$

20.3.5.3 Triple-layer coatings for substrates with low refractive index. In this subsection, the term low index substrate is used in the same sense as in 20.3.3.2.1. For such substrates, a quarter-half-quarter coating is quite effective in reducing the reflectivity to below .005 throughout the entire visible spectral region. The computed spectral reflectivity of such a coating on a glass substrate of index 1.51, is shown in Figure 20.43. The reflectivity of single-layer, double-quarter, and quarter-half coatings are also shown for purposes of comparison. The reflectivity of the triple-layer coating is well below the reflectivity of the single-layer coating over an octave (i.e. from $g = 0.6$ to $g = 1.2$). However, outside of this region, the reflectivity rises to quite large values. Hass[31] presents a detailed study of this type of coating and shows how the spectral reflectivity is altered as the refractive index and the thickness of each layer is varied. Figures 20.44, 20.45, 20.46, and 20.47 show the measured spectral reflectance of glass covered with this type of triple-layer coating.

20.3.5.4 Triple-layer coatings for substrates with high refractive index. The term high refractive index is used in the same sense as in 20.3.3.4.1. For such substrates, it is possible to use coating materials whose refractive indices satisfy

$$n_o < n_1 < n_2 < n_3 < n_s \quad . \quad (37)$$

Figures 20.32 and 20.34 show the spectral reflectivity of a triple layer coating on a substrate of arsenic trisulfide glass whose refractive indices satisfy Equations (36) and (37). The reflectivity does not exceed

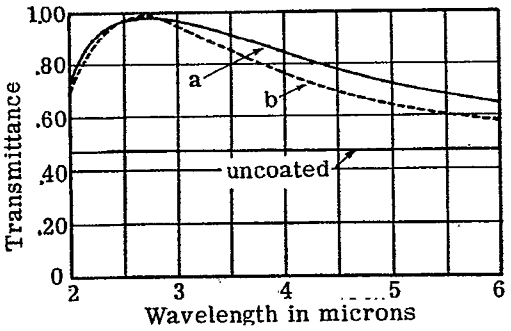


Figure 20.41 – Measured transmittance of Ge plate with antireflection coatings of (a) Si + didymium fluoride, and (b) SiO; (nt = λ/4 at 2.7 μ). From ref. 28 .

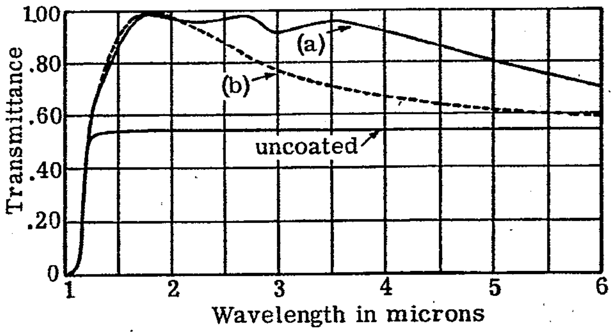


Figure 20.42 – Measured transmittance of Si plate with antireflection coatings of (a) CeO_2 + MgF_2 (nt = λ/4 at 2.2 μ), and (b) SiO (nt = λ/4 at 1.8 μ). From ref. 28.

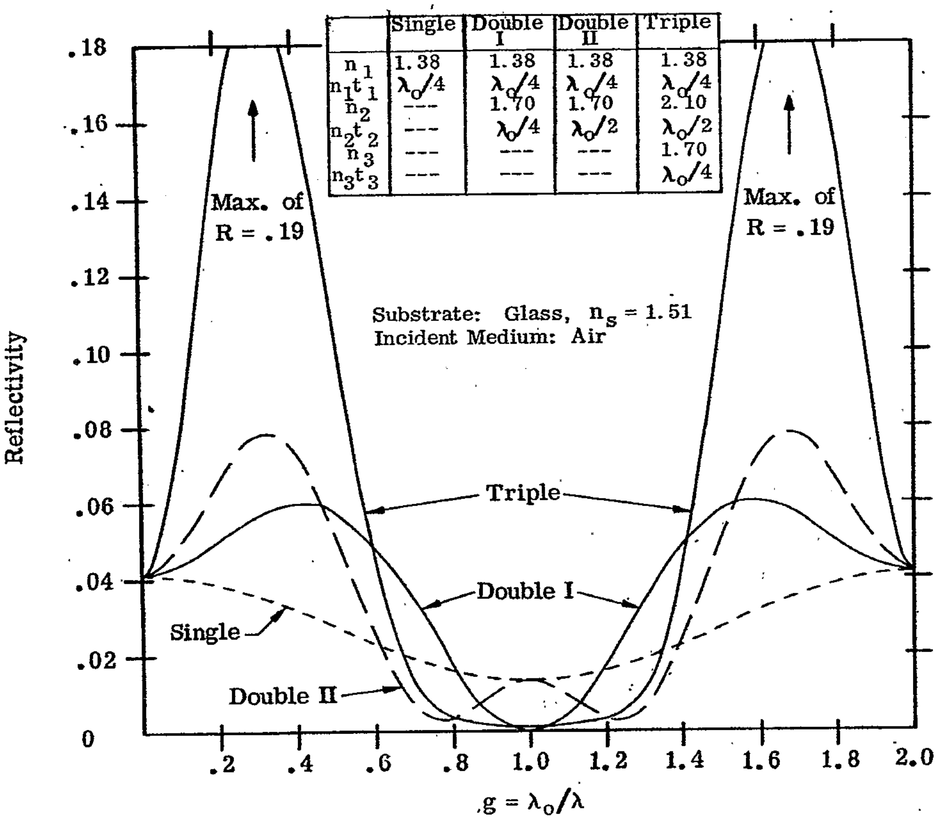


Figure 20.43 – Computed spectral reflectivity of antireflection coatings at normal incidence.

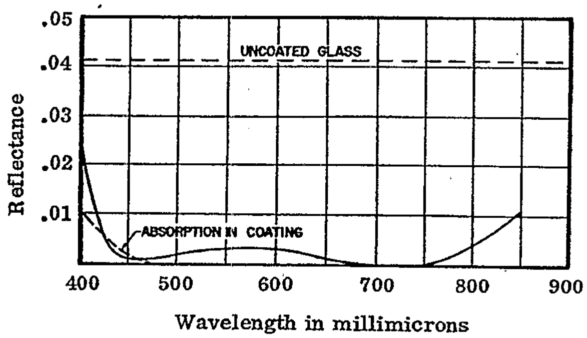


Figure 20.44 - Measured reflectance of a quarter-half-quarter antireflection coating consisting of $MgF_2 + ZrO_2 + CeF_3$ on glass. From ref. 31.

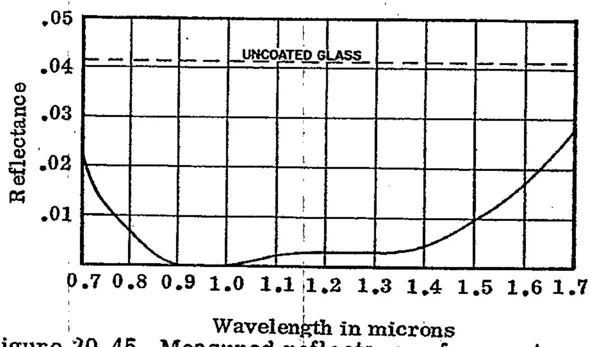


Figure 20.45 - Measured reflectance of a quarter-half-quarter antireflection coating consisting of $MgF_2 + Nd_2O_3 + CeF_3$ on glass. From ref. 31.

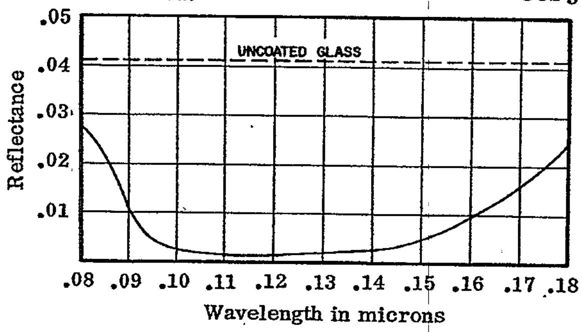


Figure 20.46 - Measured reflectance of a quarter-half-quarter antireflection coating consisting of $MgF_2 + SiO + CeF_3$ on glass. From ref. 31 .

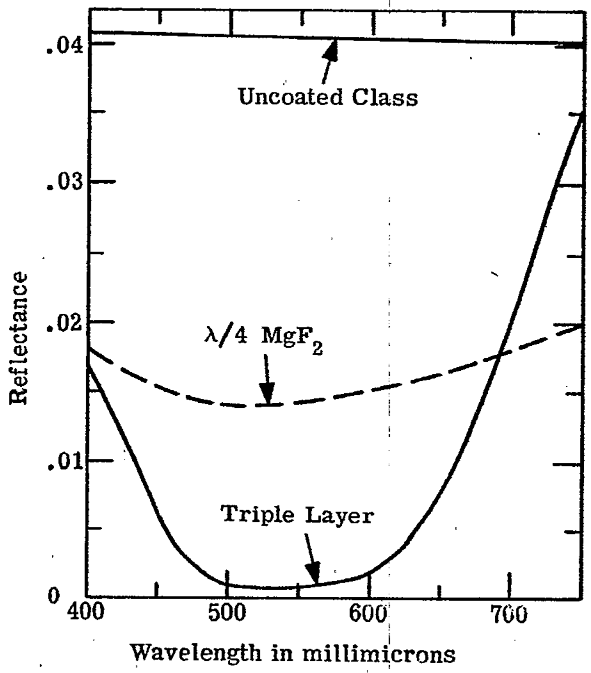


Figure 20.47 - Measured spectral reflectivity of single and triple-layer antireflection coatings. Courtesy of Fish-Schurman, Corporation.

.005 over a frequency range of more than two octaves. Silicon monoxide and chiolite could be used as films in such a coating. Figure 20.36 shows a coating of similar design on a silicon substrate.

20.4 THE REFLECTIVITY OF MULTILAYERS WITH PERIODIC STRUCTURE

20.4.1 The quarter-wave stack.

20.4.1.1 The basic period.

20.4.1.1.1 Before delving into the subject of multilayer mirrors, color filters, beam splitters, etc., it is helpful to understand some of the basic concepts relating to the propagation of light in a multilayer stack with a periodic structure.

20.4.1.1.2 The simplest type of stack with a periodic structure is a quarter-wave stack, which consists of layers with the same optical thickness, but alternating between two refractive indices, n_a and n_b. A diagram of a quarter-wave stack consisting of eight layers is shown in Figure 20.48. At some wavelength λ_o, the optical thickness of each layer is $\lambda_o/4$, that is

$$n_i \; t_i = \lambda_o/4 \tag{38}$$

for all values of i. The basic period of the quarter wave stack consists of two layers: HL, using the notation of Section 20.1.3.5. The design of the stack depicted in Figure 20.48 can be written:

Glass LHLHLHLH Air .

The foregoing design can also be written:

Glass $(LH)^4$ Air

or as

Glass $(LH)^m$ Air, where $m = 4$,

to emphasize the fact the basic period, L H, is repeated "m" times.

20.4.1.2 Reflectivity of a typical quarter-wave stack. Figures 20.49 and 20.50 show the computed reflectivity versus $g = \lambda_o/\lambda$ of a quarter-wave stack with the same n_a and n_b but with different numbers of periods. The following observations are made:

(1) The reflectivity in the range of g within the crosshatched area monotonically increases, as the number of basic periods, m, increases from 2 to 5. This region is called the high-reflectance zone.

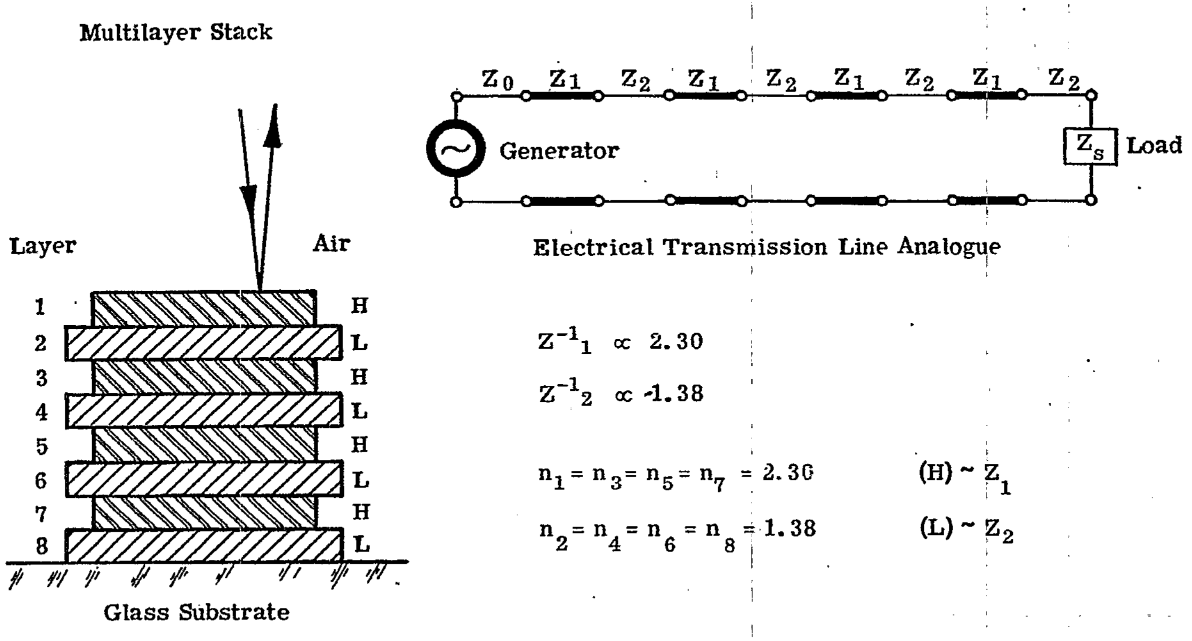
(2) The reflectivity outside of the high reflectance zone is an oscillating function. By this we mean that for any arbitrary value of g (outside of the high-reflectance zone) the reflectivity can either increase or decrease if two addition layers (i.e. L H) are added to the stack so that m increases to $m + 1$.

(3) In the region between $g = 0$ and the edge of the high-reflectance zone, there are $m - 1$ maximum and $m - 1$ minimum in the reflectivity.

20.4.1.3 Properties of an infinite stack. It is of interest to consider the limiting case where the number of periods, m, becomes infinite. This is called the infinite stack. The statements in 20.4.1.2 can be generalized to include any quarter-wave stack:

(1) The reflectivity in the high-reflectance zone approaches 1.0 as m becomes infinite. As the number of periods becomes large, the reflectivity curve in the high-reflectance zone becomes very flat. Regardless of the number of periods the reflectivity attains a maximum value R_{max} at the center of the zone at $g = 1.0$. The width of the high-reflectance zone depends only on the index ratio, n_a/n_b.

(2) The reflectivity curve in the region outside of the high-reflectance zone oscillates between maximum and minimum values. The number of oscillations between $g = 0$ and the edge of the high-reflectance zone depends upon the relationship between n_o, n_s, n_a, and n_b, but is proportional to m for $m >> 1$.



$Z^{-1}{}_1 \propto 2.30$

$Z^{-1}{}_2 \propto 1.38$

$n_1 = n_3 = n_5 = n_7 = 2.30$ (H) ~ Z_1

$n_2 = n_4 = n_6 = n_8 = 1.38$ (L) ~ Z_2

Figure 20.48 - Diagram of a quarter-wave stack and its transmission line analogue.

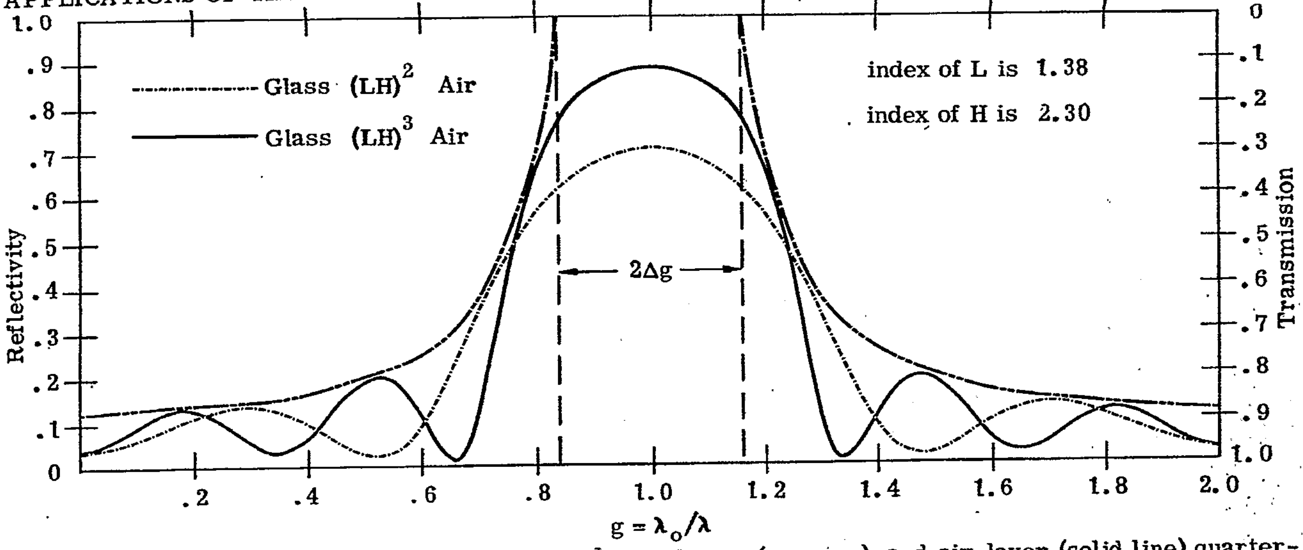


Figure 20.49 - Computed spectral reflectivity of a four-layer (-·-·-·-) and six-layer (solid line) quarter-wave stack and the envelope of maximum reflectivity (-----------).

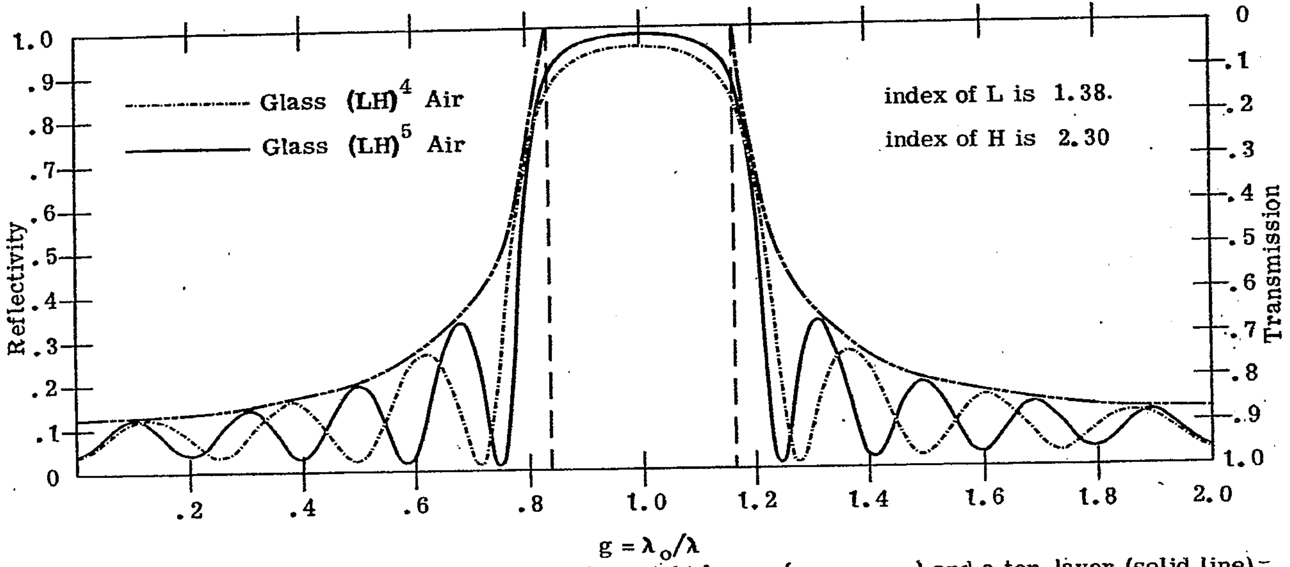


Figure 20.50 - Computed spectral reflectivity of an eight layer (-·-·-·-·-) and a ten-layer (solid line) quarter-wave stack, and the envelope of maximum reflectivity (-----------).

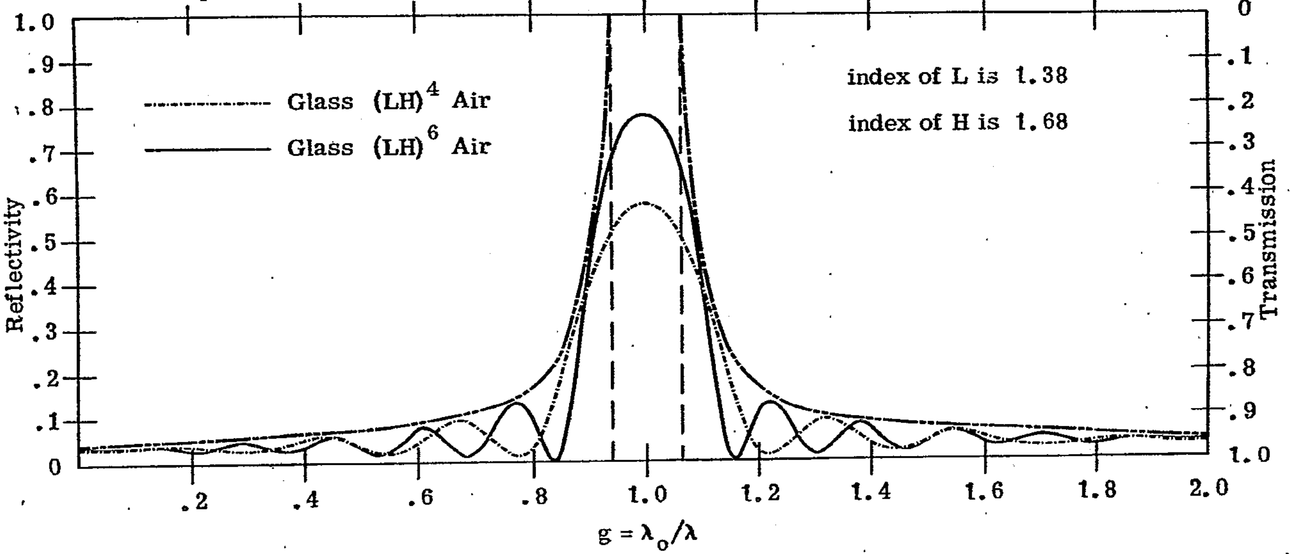


Figure 20.51 - Computed spectral reflectivity of an eight-layer (-·-·-·-) and a twelve-layer (solid line) quarter-wave stack, and the envelope of maximum reflectivity (-----------).

The reflectivity oscillates between two envelope curves. The reflectivity maximum envelope curve is shown as dashed line - - - - - - - in Figures 20.49 to 20.52. Both the maximum and minimum envelope curves are shown in Figure 20.53.

(3) The reflectivity of a multilayer with a periodic structure is adequately described by plotting the width of the high-reflectance zone and the maximum and minimum reflectivity envelopes, as is shown in Figure 20.54.

20.4.1.4 Width of the high-reflectance zone. The high-reflectance zone is symmetrical about $g = 1.0$. In the case of a quarter-wave stack, there is a simple expression for the distance Δg from the center of high-reflectance zone at $g = 1.0$ to its edge:

$$\Delta g = \frac{2}{\pi} \arcsin \left| \frac{1 - n_a/n_b}{1 + n_a/n_b} \right| \tag{39}$$

where the $| \quad |$ denotes an absolute value; the principal value of the arcsin is to be used. The total width of the high-reflectance zone is $2\Delta g$. For example, the quarter stacks shown in Figures 20.49 and 20.50 have $n_a = 1.38$ and $n_b = 2.30$; from Equation (39) we find that $\Delta g = 0.161$. As another example, consider the quarter-wave stacks shown in Figures 20.51 and 20.52. The index ratio, n_a/n_b of these stacks is approximately the same, although the stack in Figure 20.51 is composed of low-index layers, while the latter stack has high-index layers. As one might expect, the width of the high-reflectance zone of the two stacks is essentially equal. The index ratio n_a/n_b is large for the stack shown in Figure 20.53 and consequently the high-reflectance zone is quite wide. This large index ratio can only be attained with materials such as germanium and chiolite in the infrared.

20.4.1.5 Maximum reflectivity of a quarter-wave stack. Consider a more general type of quarter-wave stack in which the optical thickness of each of the layers is $\lambda_o/4$, but the refractive index, n_i, of each of the layers can be different. From Equations (5), (7), (9) and (10), it follows that the transmission T_{min} and reflectivity R_{max} at $g = 1.0, 3.0, 5.0$, etc. is

$$R_{max} = 1 - T_{min} = (P + P^{-1} - 2)(P + P^{-1} + 2)^{-1} \tag{40}$$

$$T_{min} = 4/(P + P^{-1} + 2) \tag{41}$$

where the variable P is defined as

$$P = \left[\frac{n_\ell}{n_{\ell-1}} \; \frac{n_{\ell-2}}{n_{\ell-3}} \cdots \frac{n_4}{n_3} \; \frac{n_2}{n_1} \right]^2 \frac{n_o}{n_s} \tag{42}$$

when the total number of layers ℓ is even and

$$P = \left[\frac{n_\ell}{n_{\ell-1}} \cdot \frac{n_{\ell-2}}{n_{\ell-3}} \cdots \frac{n_3}{n_2} \; n_1 \right]^2 \frac{1}{n_o n_s} \tag{43}$$

when ℓ is odd. From Equation (41) we see that as the number of layers becomes large, P also becomes large and $p >> p^{-1}$, $p >> 2$ and thus

$$T_{min} \sim 4P^{-1} . \tag{44}$$

Equations 40 to 43 are for a quite general type of quarter-wave stack, and can be easily applied to specific cases. For example, P for the stack in Figure 20.48 is:

$$P = \left[\frac{n_b}{n_a} \right]^8 \frac{n_o}{n_s} . \tag{45}$$

In the case where the stack has many layers and hence the index ratio n_a/n_b is raised to a large power, T_{min} is a very sensitive function of this ratio.

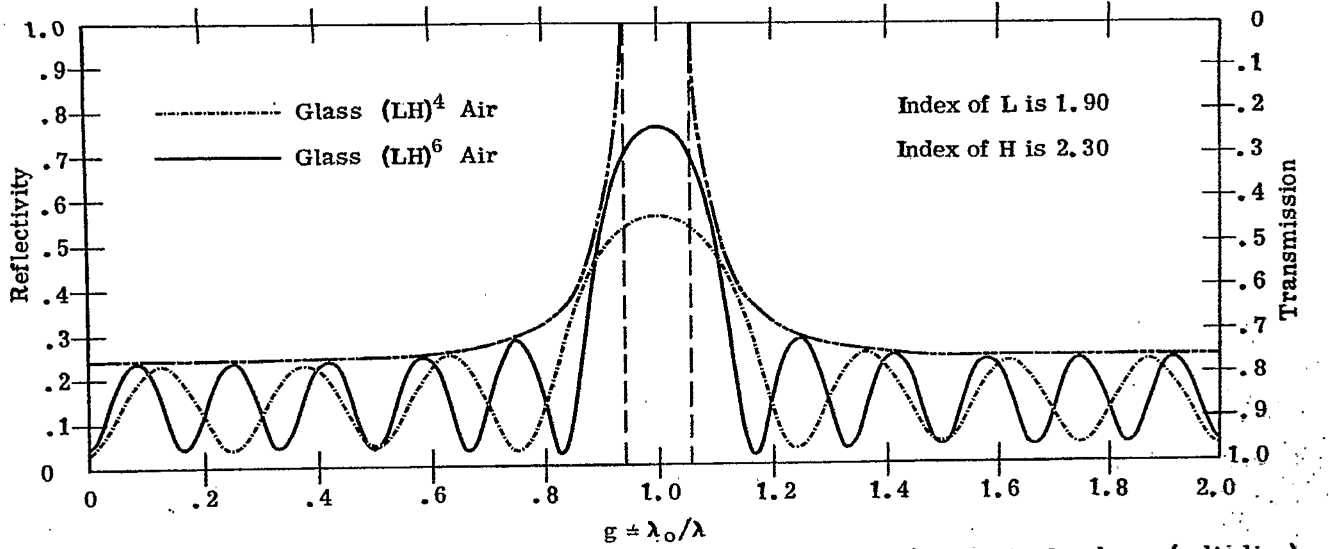


Figure 20.52 - Computed spectral reflectivity of an eight-layer (—·—·—·—) and a twelve-layer (solid line) quarter-wave stack, and the envelope of maximum reflectivity (—·—·—·—).

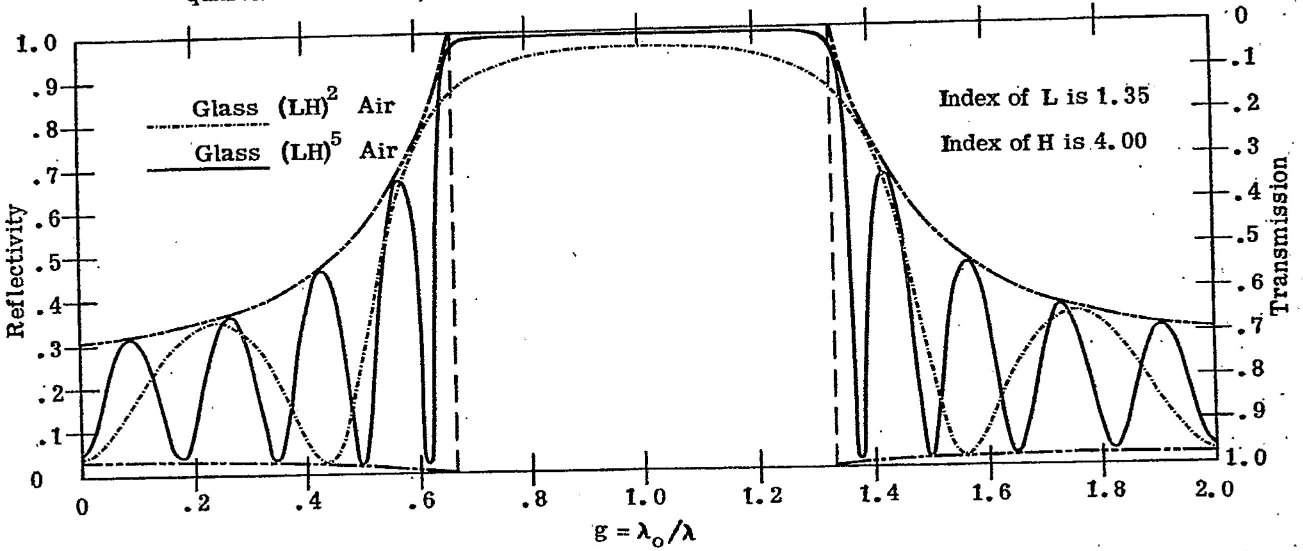


Figure 20.53 - Computed spectral reflectivity of a four-layer (—·—·—·—) and a ten-layer (solid line) quarter-wave stack, and envelope of maximum reflectivity (—·—·—·—).

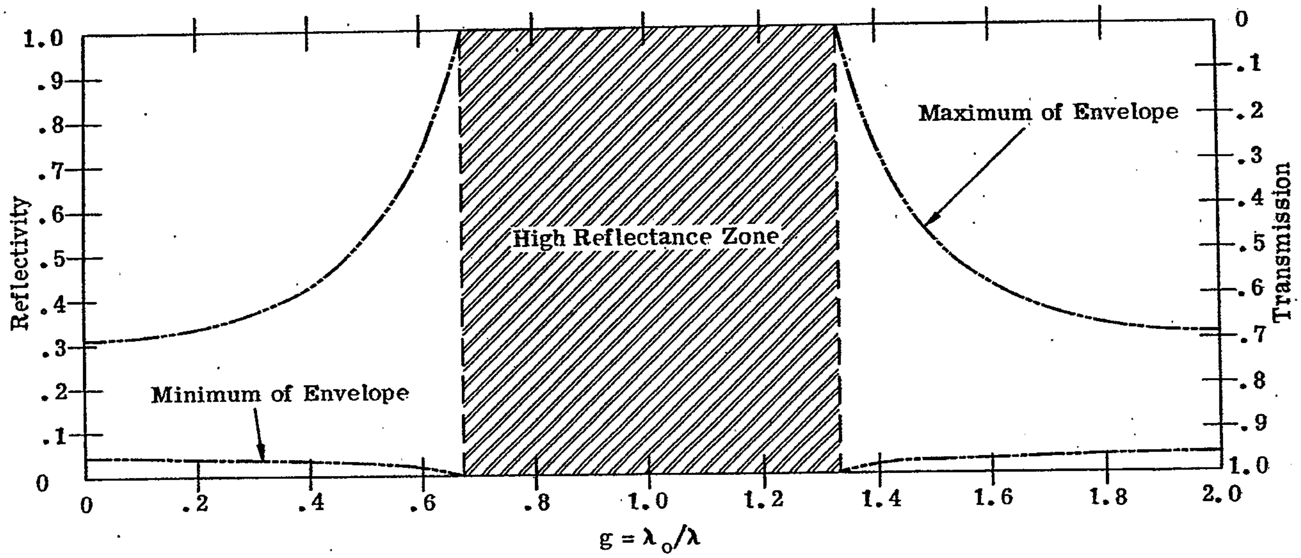


Figure 20.54 - Showing the high-reflectance zone, and the minimum and maximum of the reflectivity envelope of the quarter-wave stacks shown in Fig. 20.53

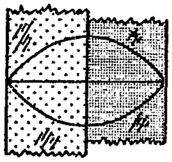
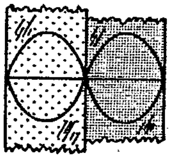
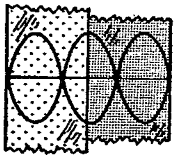
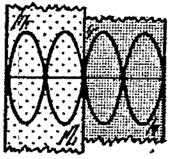
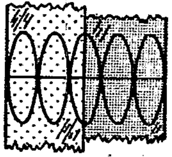
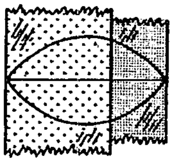
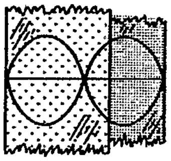
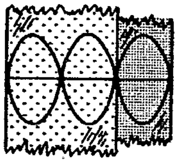
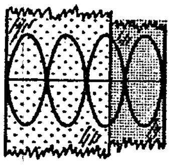
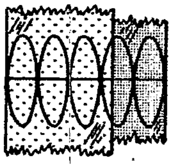
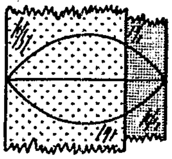
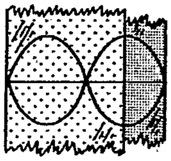
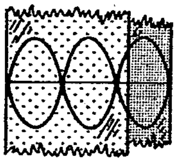
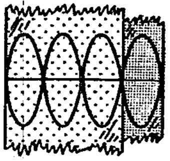
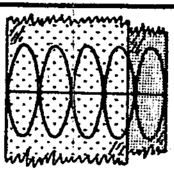
Type of Stack	g = 1.0	g = 2.0	g = 3.0	g = 4.0	g = 5.0	R_{max}, First Order, m=4
Quarterwave Basic Period L H						.957
2 : 1 Basic Period L' L' H'						.925
3 : 1 Basic Period L" L" L" H"						.858

Figure 20.55 A comparison of the wavelength (in vacuo) with the optical thickness of a basic period of a multilayer with a periodic structure, for various orders of interference.

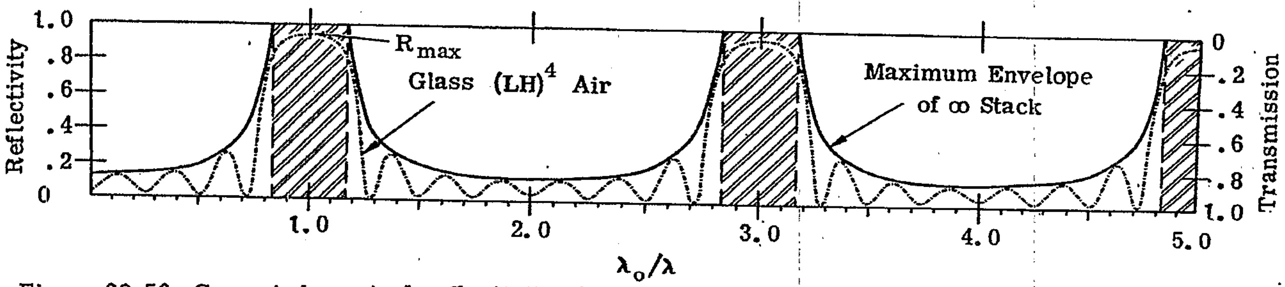


Figure 20.56 - Computed spectral reflectivity of an eight-layer quarter-wave stack (------------) and its envelope of maximum reflectivity (solid line). $n_H t_H + n_L t_L = \lambda_o/2$

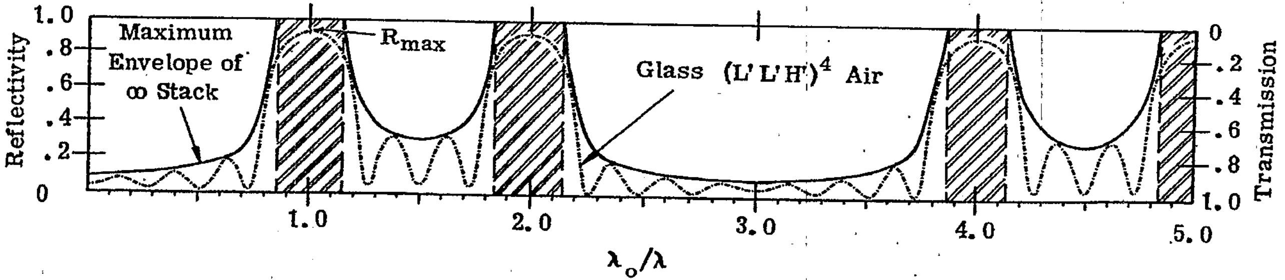


Figure 20.57 - Computed spectral reflectivity of an eight-layer 2:1 stack (------------) and its envelope of maximum reflectivity (solid line). $2\, n_L t_L + n_H t_H = \lambda_o/2$

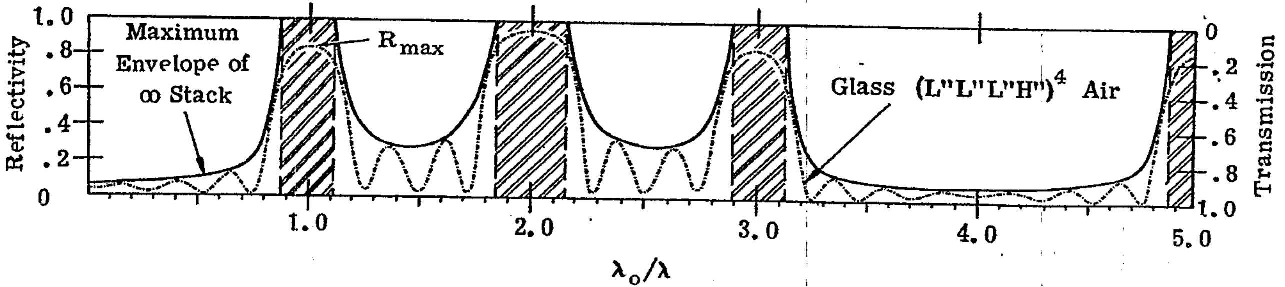


Figure 20.58 - Computed spectral reflectivity of an eight-layer 3:1 stack (------------) and its envelope of maximum reflectivity (solid line). $3\, n_L t_L + n_H t_H = \lambda_o/2$

20.4.2 Stacks with unequal thickness ratios.

20.4.2.1 General analysis.

20.4.2.1.1 Thus far we have considered the reflectivity properties of only a very specialized type of multilayer with a periodic structure, namely the quarter-wave stack, in which the two layers in a basic period have equal optical thickness. The high-reflectance zone occurs when the optical thickness of each layer is $\lambda_o/4$. Another way of stating this is to say that the high-reflectance zone occurs when the optical thickness of an entire period, LH equals $\lambda_o/2$. As is shown schematically in Figure 20.55, the high-reflectance zone occurs when a half-wave length fits into a basic period of stack. Consider the more general case where the two layers which compose the basic period of the stack do not have equal optical thickness, as for example, the stack

glass $(HL')^6$ air

where the optical thickness of the L' layer is arbitrarily chosen to be 23% greater than the H layer. Does such a stack possess a high-reflectance zone ? The answer is yes. The position of the high-reflectance zones in any multilayer with a periodic structure can be found with the aid of the rule stated in the following paragraph.

20.4.2.1.2 A necessary but not sufficient condition for a high-reflectance zone to occur at a wavelength λ_o in a stack with a periodic structure is

$$\sum_i n_i t_i = q \frac{\lambda_o}{2} \qquad (46)$$

where q is an integer and the summation is over the basic period of the stack. Stated in other terms, the sum of the optical thicknesses of the layers comprising a basic period should equal an integral number of half wavelengths. This fact and the concept of an absentee layer (see Section 20.1.5.2.2) enables one to determine where the pass bands and high-reflectance zones of multilayer with a periodic structure occur.

20.4.3 Quarter-wave stack. Figure 20.56 shows the envelope function of the same quarter-wave stack shown in Figure 20.50, but over a larger range of g. Equation (46) is satisfied at $g = 1.0$ and hence this is in a high-reflectance zone. When $g = 2.0$, the optical thickness of the basic period LH is two half-waves and Equation (46) is satisfied for $q = 2$. Hence a high-reflectance zone could exist, but does not, because each of the layers in the stack is an absentee layer. Thus the reflectivity at $g = 2.0$ is the same as that of uncoated glass, namely .041. When $g = 3.0$, the total optical thickness of the basic period LH is three half-waves, and another high-reflectance zone occurs. When $g = 4.0$, the optical thickness of each layer is two half-waves and consequently all of the layers are absentee. Another high-reflectance zone occurs at $g = 5.0$, when the total optical thickness of the basic period LH is five-half waves. In the quarter-wave stack, high-refeitance zones occur at $g = 1, 3, 5, 7, 9, \ldots$. that is at odd integers. The high-reflectance zone which occurs at $g = 3$ is the third harmonic of the high-reflectance zone which occurs at $g = 1$, or to use the terminology of physical optics, this is a third-order interference peak. This is analogous to an open-ended organ pipe, which can sustain only odd harmonics.

20.4.3.1 The 2:1 stack. Let us study the reflectivity properties of the 2:1 stack:

glass L'L'H' L'L'H' L'L'H' L'L'H' L'L'H' etc.

which can also be written as

glass $(L'L'H')^m$ air

where m is an integer. Here we have used the primed superscript, i.e. L' and H', to show that the optical thickness of these layers is different from the L and H which were used in the quarter-wave stack described in Section 20.4.1. In both cases, the optical thickness of the layers has been chosen so that a first-order interference high-reflectance zone occurs at λ_o. One must remember that the combination L'L' represents a single layer because each L' layer has the same refractive index. Thus the optical thickness of the L'L' layer is twice that of the H' layer and hence this is called a 2:1 stack. The spectral reflectivity curve of a 2:1 stack and its maximum reflectivity envelope are shown in Figure 20.57. Applying the rule which was stated in Section 20.4.2.1, the first-order high-reflectance zone occurs at $g = 1.0$ when the total optical thickness of the basic period L'L'H' is $\lambda_o/2$, as is shown in Figure 20.55. The second-order high-reflectance zone occurs at $g = 2.0$ when the total optical thickness of the basic period is two half-waves of λ_o. One might expect another high-reflectance zone at $g = 3.0$, but the optical thickness of the H' layer is a half-wave and the optical thickness of the L'L' layer is two-half waves. Hence, at $g = 3.0$, all of the layers are absentee layers and the reflectivity is that of uncoated glass. At $g = 4.0$ and $g = 5.0$, the total optical thickness of the basic period equals four half-waves and five half-waves, respectively and high-reflectance zones occur. In this example, we have chosen $L'L'H'$ as a basic period in which the optical thickness of the low-index layer is twice that of the high-index layer. The width of the high-reflectance zone is the same as the 2:1 stack, $H'H'L'$. As is shown in 20.4.6 however, the reflectivity properties at non-normal incidence of the two types of 2:1 stacks are quite different.

20.4.3.2 The 3:1 stack.

20.4.3.2.1 As a final illustration, consider the properties of stack with a periodic structure in which the optical thickness ratio is 3:1 :

$$\text{glass } (L''L''L''H'')^m \text{ air}.$$

The double prime superscripts, H'' and L'' are used to show that the optical thickness of the L'' and H'' layers are different from both the H and L layers in the quarter-wave stack and the H' and L' layers of the 2:1 stack. The optical thickness of the H'' and the L'' layers is chosen so that the total optical thickness of the basic period of the stack, $L''L''L''H''$ equals $\lambda_o/2$, as is shown in Figure 20.55. Figure 20.58 shows a spectral reflectivity curve and the reflectivity envelope of a 3:1 stack. A first-order high-reflectance zone occurs at $g = 1.0$ and high-reflectance zones also occur at $g = 2.0$ and $g = 3.0$ when the total optical thickness of the basic period is two and three half-waves of λ_o, respectively. However, when $g = 4.0$ the optical thickness of the $L''L''L''$ layer is three half-waves and the optical thickness of the H' layer is a single half-wave. Hence all of the layers are absentee layers and the reflectivity at $g = 4.0$ is the same as uncoated glass. At $g = 5.0$ the total optical thickness of the basic period is five half-waves and hence another high-reflectance zone occurs. Thus every fourth high-reflectance zone is missing. Another way of stating this is to say that high-reflectance zones occur at the 1, 2, 3, 5, 6, 7, 9, 10, 11th etc. harmonics of the frequency of the fundamental.

20.4.3.2.2 The reflectivity curves of the multilayers with a periodic structure have been plotted on a frequency scale. It is possible that many readers are more accustomed to thinking in terms of wavelength, and so in Figure 20.59 is depicted the reflectivity versus wavelength of the 3:1 shown in Figure 20.58 with λ_o chosen to be 2.0 μ. A first-order reflectivity peak occurs at the fundamental wavelength, 2.0 μ. A second-order peak occurs at 1.0μ, which is one-half the fundamental and a third-order peak occurs at 0.667 μ, which is one-third of the fundamental. A fourth-order peak does not occur at $2.0/4 = 0.5\ \mu$ because the layers are absentee. A fifth order peak occurs at $\lambda_o/5 = 0.400\ \mu$, and so on. In comparing Figures 20.58 and 20.59, we note that on a frequency scale, the width of the first and third order reflectivity peaks is the same, whereas this is not true on a wavelength scale.

20.4.4 The general p:q stack.

20.4.4.1 General properties. It is patent that the analysis which we have made on the properties of the 1:1 (i.e. quarter-wave stack), 2:1 and 3:1 stack could be easily extended to any stack in which the ratio of the optical thickness of the layers is p:q, where q and p are integers. The following comparisons can be made between stacks which have a periodic structure, but different thickness ratios:

(1) The high-reflectance zone of the quarter-wave stack has even symmetry about $g = 1.0$, 3.0 etc. This is not necessarily true for other types of stacks.

(2) For a stack composed of layers of alternating refractive index n_a and n_b the width of the high-reflectance zone for any given ratio of $n_a : n_b$ is the largest when the optical thickness is equal. In other words, the high-reflectance zone of the quarter-wave stack is wider than high-reflectance zone of 2:1, 3:1, 3:2, etc. stacks.

(3) The width of the high-reflectance zone of a quarter-wave stack is given by a rather simple expression (i.e. Equation (39)). No such simple equations exist for other types of stacks. For example, the width of the high-reflectance zone of a 2:1 stack is given by the roots of a cubic equation.

(4) In the spectral region outside of the high-reflectance zones, the number of oscillations of the reflectivity increases as the number of periods (and hence the number of layers) of the stack increases. In any case, the oscillations will lie between the maximum and minimum reflectivity envelope of the infinite stack. The shape of this envelope function depends upon:

(a) The ratio of the optical thickness of the layers in a basic period.

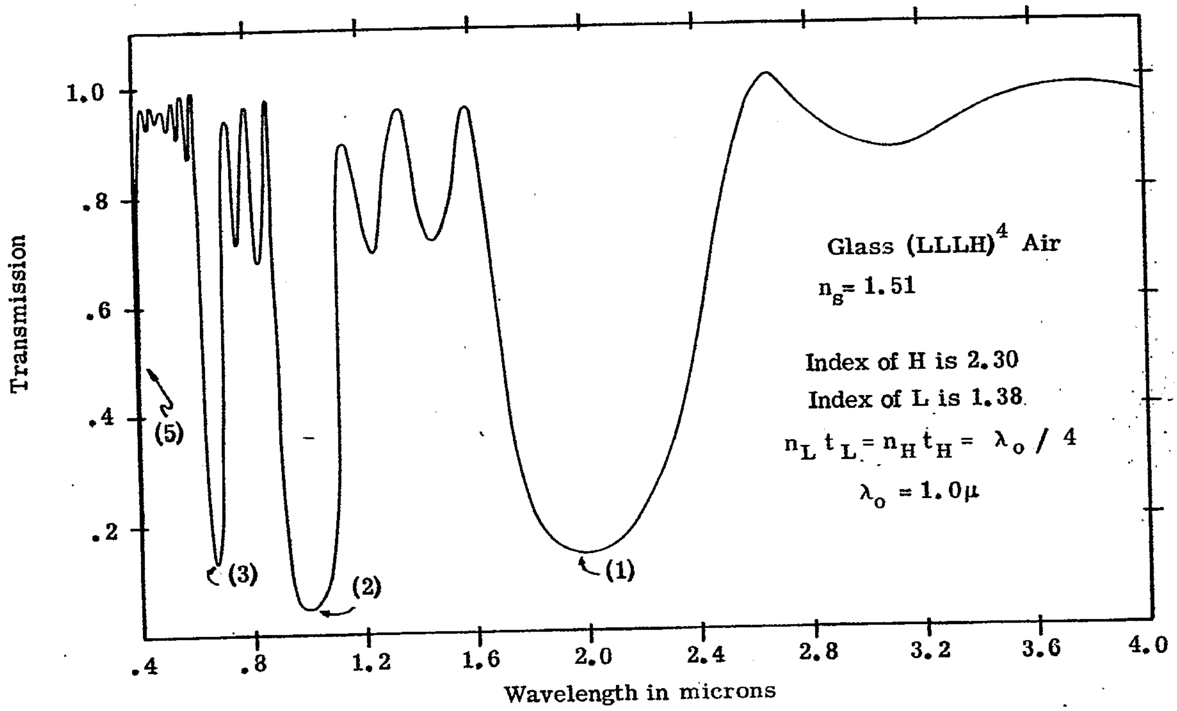


Figure 20.59 - Computed spectral transmission of an eight-layer 3:1 stack.

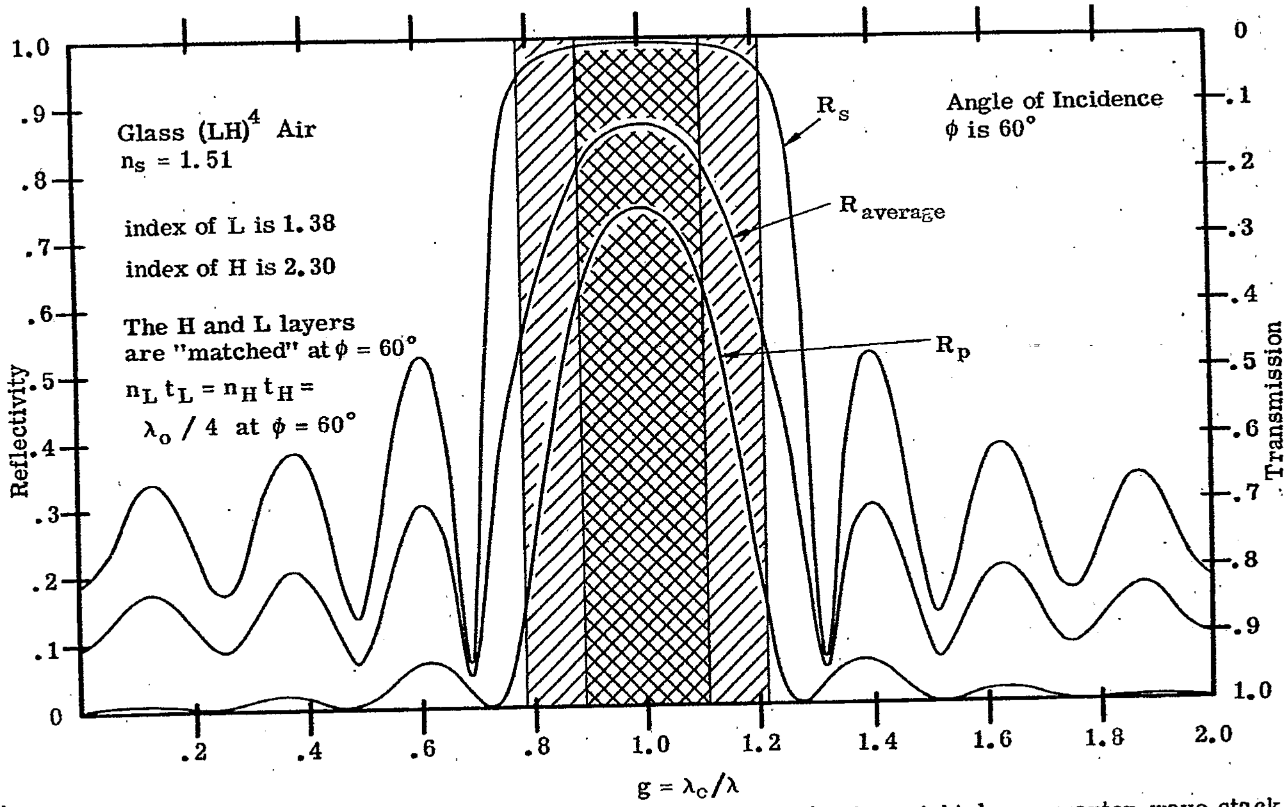


Figure 20.60 - Computed R_p, R_s, and $R_{av} = 1/2\ (R_p + R_s)$ of an eight-layer quarter-wave stack at $\phi = 60°$. The optical thickness of the H and L layers is matched 1:1 at $\phi = 60°$.

(b) The index of the two layers, n_a and n_b .

(c) The refractive index of the substrate, incident medium, and the refractive index and optical thickness of any layers which are added to the either end of the basic stack. This topic is discussed in Section 20.4.8.

(5) Given the indices n_o , n_s , n_a , n_b for any given order of interference the highest reflectivity for a given number of layers is obtained with a quarter-wave stack. This is illustrated in Figure 20.55, where the maximum reflectivity in the first order is listed for the quarter-wave, 2:1 and 3:1 stacks with an equal number of layers, with the same n_a, n_b, n_o, and n_s.

20.4.4.2 As an example of how the concept of absentee layers can be used to find the reflectivity at specific wavelengths, consider the ten-layer stack:

glass H'H'L' H'H'L' H'H'L' H'H'L' H'H'L' air

in which the optical thickness of the H'H' and L' layers has been chosen so that a first-order high-reflectance zone occurs at $g = 1.0$ as is shown in Figure 20.55. The problem is to find the reflectivity of this stack at $g = 1.5$. The optical thickness of each of the H'H' layers at $g = 1.5$ is a half-wave and hence each of these layers is absentee. Thus the stack reduces to:

glass L' L' L' L' L' air .

However, at $g = 1.5$ each of the L' layers is a quarter-wave in optical thickness and hence each pair of L' layers represents a half-wave. Thus four of the L' layers can be removed from the stack leaving:

glass L' air .

Thus the reflectivity of this ten-layer stack at $g = 1.5$ is the same as the reflectivity of a single quarter-wave layer on glass and can be readily computed from Equation (30).

20.4.4.3 The treatment given here of the propagation of waves in a medium with a periodic structure has been descriptive and qualitative. A rigorous mathematical treatment is given by Epstein[32], Brillouin[33], and Seitz[34].

20.4.5 Analogies.

20.4.5.1 Electrical transmission line. It is helpful to give some analogies between the propagation of light in a multilayer with a periodic structure and other fields, such as electrical engineering, x-ray crystallography, and solid state physics. Using the concepts which were presented in Section 20.1.4.1, the electrical transmission line shown in Figure 20.48 is the analogue of the quarter-wave stack shown in the same Figure. The transmission line consists of eight sections, each of which has an electrical length of a quarter-wave and which alternate between a high and low impedance. We recall that the load admittance of the transmission line is analogous to the refractive index of the substrate. As one can readily show by using a Smith chart, the load impedance is reflected back to the terminals of the preceeding section as a maximum value when the electrical length of the line is a quarter-wave. In this case the substitutional impedance at the input terminals is very large. Consequently there is large impedance mismatch between the characteristic impedance of the generator (which is 1.0) and the substitutional impedance of the line. This means that the voltage standing wave ratio is large and hence the voltage reflection coefficient to close to one.

20.4.5.2 Bragg reflection of x-rays. A beam of x-rays travels through a vacuum with the velocity of light. This velocity is pertubed slightly when the beam travels through a cloud of electrons, the amount of the perturbation being nearly proportional to the density of the electrons. In a crystal, there is a high density of electrons near each atomic lattice site. Since atoms are regularly spaced in a crystal, the x-rays travel through a periodically stratified medium. Although the change in the velocity is not abrupt, as it is for light propagating through a multilayer, nevertheless Bragg reflection of the x-rays is observed when the path difference between adjacent reflecting crystal planes is an integral number of wavelengths. This point is further amplified by Brillouin.[33]

20.4.5.3 The propagation of electrons in a crystal. The propagation of a single electron in a crystal is treated quantum mechanically by solving the Schrödinger equation, in which the electron is represented by a traveling wave with a deBroglie wavelength, λ_b . The velocity of the electron is perturbed by the electrostatic repulsion of electron cloud at each lattice site and consequently the electron moves in a potential which varies periodically. Regardless whether one chooses a simple one-dimensional periodical potential of Kronig and Penney, or a more sophisticated potential computed from atomic wave functions, the electron

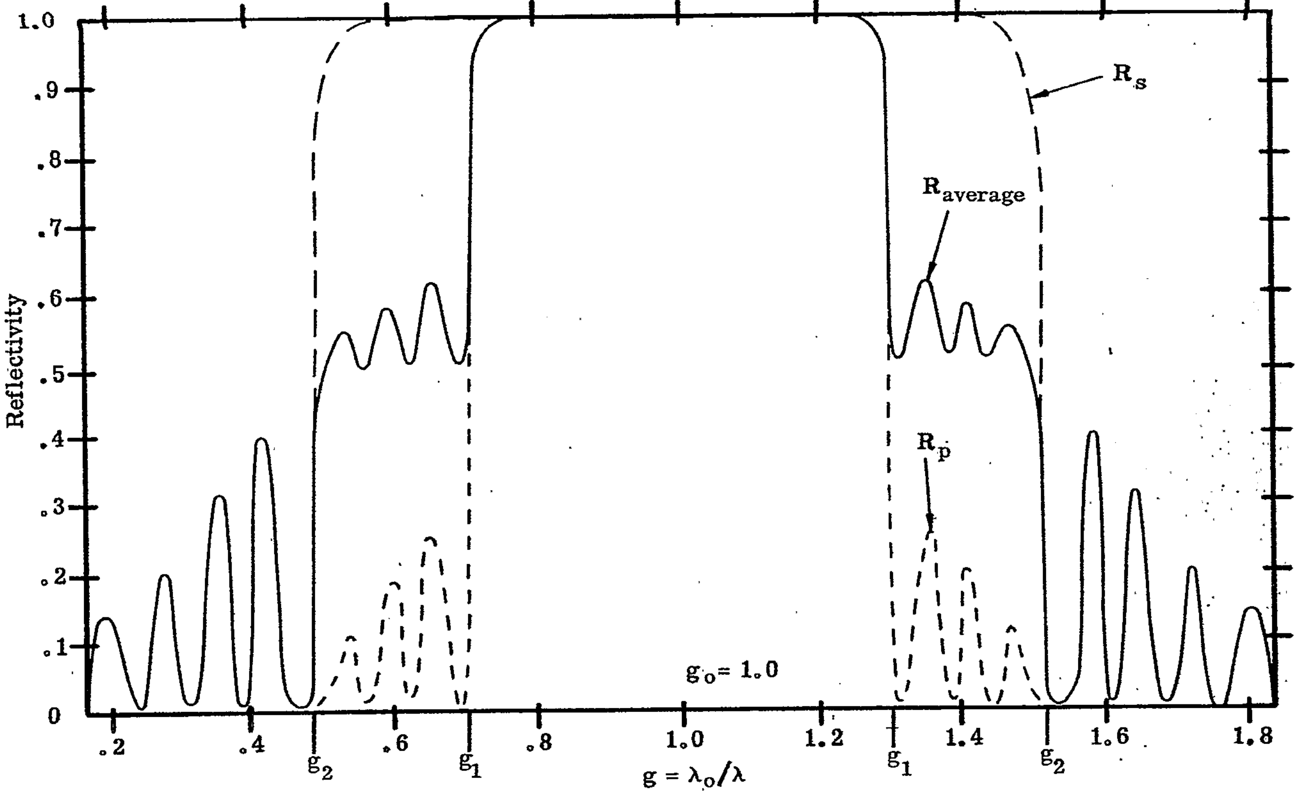


Figure 20.61 - R_p, R_s, and $R_{av} = 1/2\ (R_p + R_s)$ at $\phi = \phi_o$ of a ficticious quarter-wave stack with the optical thicknesses of the layers matched 1:1 at ϕ_o.

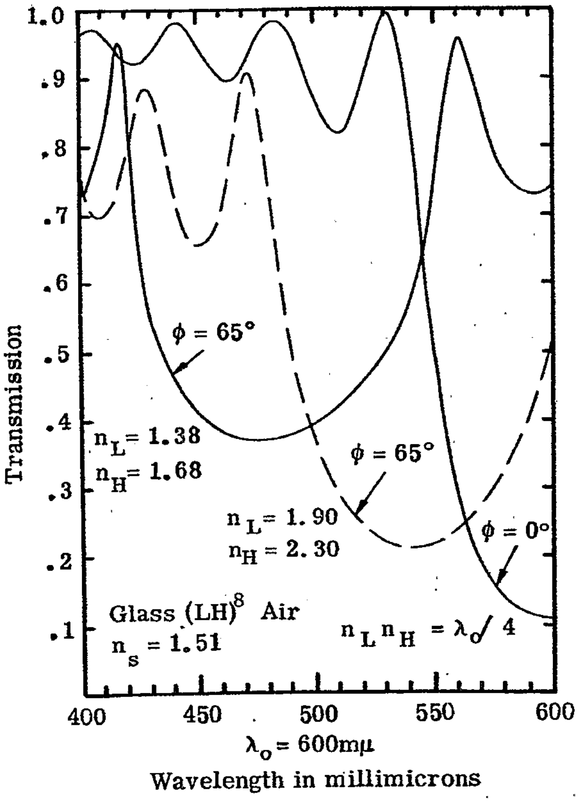


Figure 20.62 - Computed T_{av} at $\phi = 0$ and $\phi = 65°$ of sixteen-layer quarter-wave stacks with low index (solid line) and high index (dashed line).

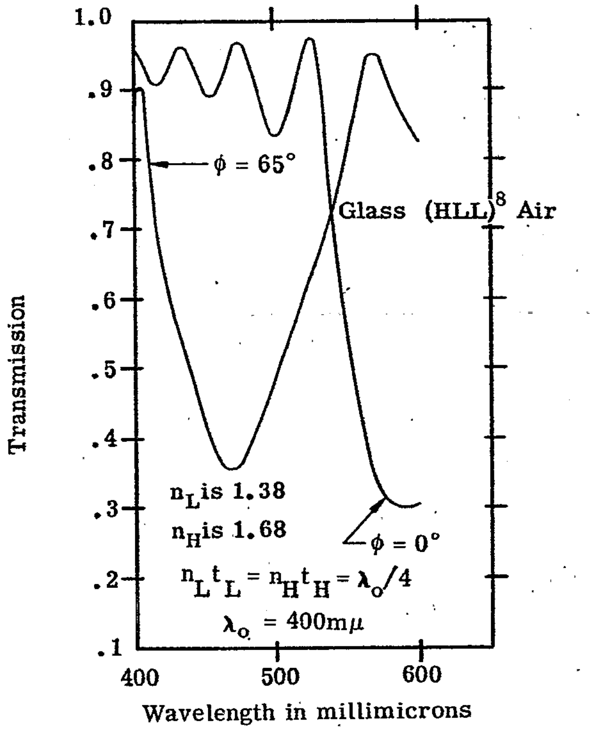


Figure 20.63 - Computed T_{av} at $\phi = 0$ and $\phi = 65°$ of a sixteen-layer 2:1 stack.

is reflected whenever $\lambda_b/2$ equals a multiple of the period of the lattice. An equivalent way of saying this is that the edge of the Brillouin zone occurs when the wave-vector $\underset{\sim}{K} = \pi/a$, where a is a lattice space in a particular direction and $\underset{\sim}{K}$ is wave-vector (Seitz)[34]. This is treated in detail by Brillouin [33], Seitz [34], and others.

20.4.6 The reflectivity of quarter-wave stacks at non-normal incidence.

20.4.6.1 Layers matched at angle. In considering the reflectivity of quarter-wave stacks at non-normal incidence, we must consider separately the case where thickness of the layers is matched (see Section 20.1.6.3, for a definition of this term) at normal incidence and the case where the layers are matched at a particular angle ϕ_o. In the latter case, the layers are deliberately mismatched so that at normal incidence the ratio of the optical thickness of the H and L layer is not 1:1, but the ratio is 1:1 at ϕ_o.

20.4.6.1.1 As an example of a match at angle, consider the eight-layer quarter-wave stack whose reflectivity curve is shown in Figure 20.60. The optical thickness of the low-index layer L has been made thicker than the high-index layer H in a ratio $n_H t_H : n_L t_L = 1.0 : 1.19$. In this case the layers are matched at $\phi_o = 60°$. Equation (39) can be used to compute the width of the high-reflection zone, using the effective index appropriate to each plane of polarization. For example, at $\phi = 60°$, in the "p" plane of polarization the effective index of the L layer is 1.773 and the effective index of the H layer is 2.483. From Equation (39) we find that the half-width, Δg, of the high-reflectance zone is 0.107. Similarly Δg in the "s" plane of polarization is 0.214, which is larger than the Δg at normal incidence, which is 0.161. The high-reflectance zones are shown in Figure 20.60 as cross hatched areas. The generalizations stated in 20.4.6.1.1 and 20.4.6.1.2 apply to quarter-wave stacks which matched at a particular angle ϕ_o, and are not confined to the eight-layer stack which has been used as an example.

20.4.6.1.2 The width of the high-reflectance zone increases in the "s" plane of polarization and decreases in the "p" plane of polarization. If many periods, say thirty or forty, are used in a stack, so that both R_p and R_s are close to 1.0 within their high-reflectance zones, the average reflectivity, $R_{av} = 1/2 (R_p + R_s)$ has the shape which is depicted in Figure 20.61. In the region between $g = 1.0$ at the center of the stack and g_1, a high-reflectance zone exists for both planes of polarization and hence R_{av} is close to 1.0. At g_1, the high-reflectance zone for the "p" plane of polarization ends and R_p fluctuates at low values outside of this zone. The high-reflectance zone for the "s" polarization extends to g_2 and in this intermediate region between g_1 and g_2 R_{av} attains a minimum value of 0.50. We observe that this shoulder on the R_{av} curve is due to the dissimilar width of the high-reflectance zones in the two planes of polarization. This shoulder has been somewhat exaggerated for purposes of illustration in Figure 20.61. The reflectivity curve of the eight-layer quarter-wave stack shown in Figure 20.60 does not show this shoulder, because the reflectance R_p does not drop to zero rapidly enough outside of the high-reflectance zone. If more periods were added to the stack, then eventually the R_{av} curve would show such a shoulder.

20.4.6.1.3 Maximum reflectivity. Since the thickness of the layers is matched at ϕ_o, the maximum reflectivity at $g = 1.0, 3.0, 5.0$, etc. can be computed in each plane of polarization from Equations (40) to (43), using the effective index appropriate to each plane of polarization.

20.4.6.2 Layers matched at normal incidence.

20.4.6.2.1 If a stack which is matched 1:1 at normal incidence is viewed at non-normal incidence, the position of the high-reflectance zone can be found by substituting effective thicknesses in Equation (46). The high-reflectance zone is no longer centered at a wavelength computed from Equation (46) and also the width of the zone (in each plane of polarization) is different than at normal incidence.

20.4.6.2.2 As an example of the application of Equation (46) to a stack at non-normal incidence, consider the 16 layer stack:

$$\text{glass} \quad (LH)^8 \quad \text{air}$$

in which both the L and H have a QWOT of 600 mμ at normal incidence, that is,

$$n_i \, t_i = 150 \text{ m}\mu = 600 \text{ m}\mu/4 .$$

The transmission curve, shown in Figure 20.62, has a minimum at 600 mμ which is the center of the high-reflectance zone. Two cases are considered separately at $\phi = 65°$.

Case I $n_L = 1.90$, $n_H = 2.30$. From Figure 20.8 we find that the change in effective thickness is 0.92 for the H layer and 0.891 for the L layer, the latter value being found by linear interpolation.

Substituting these values for the effective thickness into Equation (46), we obtain

$$(.92)\ (150\ m\mu)\ +\ (.879)\ (150\ m\mu)\ =\ \frac{\lambda_1}{2}\ , \qquad (47)$$

whence $\lambda_1 = 540\ m\mu$. From Figure 20.62 we see that the minimum T_{av} at $\phi_o = 65°$ is very close to 540 mμ .

Case II $n_L = 1.38$ and $n_H = 1.68$. The change in the effective thickness is found from Figure 20.9 to be 0.754 and 0.842. Solving an equation similar to Equation (47) gives the result that $\lambda_1 = 480\ m\mu$. Actually, Figure 20.62 shows that the wavelength at which T_{av} is a minimum is shifted to shorter wavelengths by about 10 mμ .

20.4.7 Minimization of the angle shift.

20.4.7.1 The basic problem. The transmission curves of all multilayer filters change with the angle of incidence and usually exhibit a shift to shorter-wavelengths. This angle shift is of little importance if the multilayer filter is illuminated with collimated light at one angle of incidence. However, serious problems often arise when the light is highly convergent. For example, either of the multilayers shown in Figure 20.62 would attenuate the sodium yellow lines (at 589 mμ) if used at normal incidence, but they would be quite ineffective at $\phi = 65°$. The problem of the angle shift of a multilayer filter is similar to the problem of the chromatic aberration of a lens. In either case, the effect is ubiquitous and at best the designer can only minimize it. Two methods of minimizing the angle shift are: (1) The use of high-index materials. (2) The use of more of the higher index material in the basic period.

20.4.7.2 The use of high-index materials. From the examples in 20.4.6.2.2, it is patent that the change in the effective thickness of each of the layers is much less for high-index layers than for low-index layers. Thus the angle shift of the quarter-wave stack with $n_L = 1.90$, $n_H = 2.30$ is much less than the stack with $n_L = 1.38$, $n_H = 1.68$, even though the width of the high-reflectance zone at normal incidence is the same because the index ratio n_L/n_H is the same. From this it follows that if a multilayer with a periodic structure is used as a "cutoff" filter - that is to pass,either in the long-wave or short-wavelengths region - it should contain high index materials if the angle shift is to be minimized. For example, an infrared filter which contains cryolite and germanium has a larger angle shift than a filter which contains zinc sulfide and germanium, because the refractive index of zinc sulfide is nearly twice as large as that of cryolite.

20.4.7.3 The use of more high-index material in a basic period. If only quarter-wave stacks were considered, then the discussion would terminate with 20.4.7.2. If, however, the optical thickness of the high-index material is different from the thickness of the low index material, such as occurs in a 2:1 or a 3:1 stack, then it is possible to reduce the angle shift below that of a quarter-wave stack (using the same materials) by using the high-index material in the thicker layer. That is to say, a $(LHH)^m$ stack has a lower angle shift than the stack $(HLL)^m$, even though the width of the high-reflectance zone at normal incidence of these two stacks is identical. Similarly, the stack $(LHHH)^m$ has a smaller angle shift than the stack $(HLLL)^m$. These facts are illustrated in Figure 20.63 which shows the spectral transmission of the 2:1 stack

glass $(H\ LL)^8$ air

at normal incidence and at 65°. In this stack, the optical thickness of the low-index layer is greater than that of the high-index layer. The 2:1 stack whose transmission curve is shown in Figure 20.64 has more high-index material in the basic period:

glass $(L\ HH)^8$ air.

In Figure 20.64 the T_{av} at 65° of the (L HH) stack (shown as solid line) is compared with the H LL stack (dashed line). Although both stacks show a considerable angle shift to the blue, the shift of the LHH stack is definitely less. The angle shift of the HLL stack is considerably greater than the comparable quarter-wave stack shown in Figure 20.62 which uses the same refractive indices. Figures 20.79, 20.86, and 20.91 show the angle shift of transmission curve of various multilayers.

20.4.8 Variations on the basic periodic structure.

20.4.8.1 General considerations. Most of the quarter-wave stacks and other multilayers with a periodic structure shown in Figures 20.48 to 20.64 have been selected primarily as illustrations and should be modified slightly if they are used as practical filters. Additional layers can be added to these stacks for either of two purposes:

(1) To increase the reflectivity in the high reflectance zone.

(2) To increase the transmission in the spectral region outside of the high-reflectance zone.

If a quarter-wave stack were used as a semi-transparent mirror coating for a Fabry-Perot interferometer, then primary objective would be to obtain a high reflectivity over a specific spectral region and little interest is paid to the transmission outside of this region. On the other hand, if a quarter-wave stack were used as a long-wave pass filter, then it is important to optimize the transmission in the long-wave region.

20.4.8.2 Increasing the reflectivity. It is evident from Equations (42) and (43) that a quarter-wave stack has a higher R_{max} when a high-index layer is next to both the substrate and the incident medium. The R_{max} of this odd-layered stack is greater than the R_{max} of an even-layered stack. This is illustrated in Figure 20.65, shows the computed reflectivity of a six-layer and a seven-layer stack which use the same n_a and n_b. Even though an additional H layer has been added to the basic stack so that its design can no longer be represented as $(HL)^m$, the seven-layer stack is still called a quarter-wave stack. It is evident that a considerable increase in the reflectivity has been achieved by the addition of the extra H layer. If an additional L layer were added so that the multilayer design is,

$$\text{glass } (HL)^4 \text{ air},$$

the R_{max} would be less that of the seven-layer stack. An increase in the maximum reflectivity is also achieved if an odd number of layers are used rather than an even number in the 2:1 and 3:1 stacks shown in Figures 20.57 and 20.58.

20.4.8.2.2 The effect on R of a mismatch in layer thickness. Closely allied to the problem of attaining the maximum reflectivity is the problem of a mismatch in the thickness of the layers. If the optical thickness of any one of the layers in a quarter-wave stack, or other type of multilayer with a periodic structure, deviates by a small amount from $\lambda_0/4$, then the reflectivity throughout the entire high-reflectance zone falls below the value which would be attained if all the layers were perfectly matched. For example, suppose that a quarter-wave stack is manufactured by evaporating each of the layers in a vacuum and that in this process errors of a random nature are made in controlling the thickness of each of the layers. Consequently, the optical thickness of each of the layers differs from $\lambda_0/4$ by a random amount. As long as these errors are not excessive, say greater than ten percent, the region of high-reflectivity about λ_0 which is characteristic of a quarter-wave stack, is still observed. However, the R_{max} of such a multilayer is not as high as it would be for a perfect stack and also the cutoff at the edge of the high-reflectance zone is not as steep. The fact that rather large errors can be made in controlling the thickness of the layers without serious detrimental effects upon the reflectivity is the factor which permits certain types of band pass multilayer filters to be manufactured with relatively crude monitoring equipment to control the thickness of the layers.

20.4.8.3 Enhancing the transmission in the spectral region outside of the high-reflectance zone. Several methods are used to enhance the transmission in the spectral region outside of the high-reflectance zone:

(1) Additional layers of non-quarter-wave optical thickness can be added to the stack. This is discussed in 20.4.8.3.1.

(2) It is possible to vary the thickness of each of the layers by a small amount so that the transmission increases outside of high-reflectance zone. Since this method improves upon, or refines, an existing multilayer design, it is called the refining method. This is covered in 20.4.8.4.

20.4.8.3.1 An effective method of enhancing the transmission in the spectral region outside of the high-reflectance zone is to add additional layers to the basic stack (which has a periodic structure). For example, one could start with the twelve-layer quarter-wave stack

$$\text{glass } (HL)^6 \text{ air}$$

and add several layers to either end of the stack:

$$\text{glass } L_1 \; H_2 \; (HL)^6 \; H_3 \; L_4 \text{ air}$$

where the additional layers L_1, H_2, H_3 L_4 are tagged with subscripts to emphasize the fact that they do not necessarily have the same thickness or refractive index as the H and L layers in the basic stack. As an example of this procedure, suppose that a quarter-wave stack is used as a short-wave pass filter which is intended to pass the blue and green, but attenuate in the yellow and red. A seven-layer stack of zinc sulfide $(n = 2.30)$ and magnesium fluoride $(n = 1.38)$ is considered:

$$\text{glass} \quad H\,L\,H\,L\,H\,L\,H \quad \text{air}\,.$$

The reflectivity versus frequency curve is shown in Figure 20.65, and has even symmetry about λ_o. The region of high reflectivity extending from $g = 0.8$ to 1.2 would attenuate the red and yellow, but this multilayer would be much more effective as a short-wave pass filter if the reflectivity peaks at $g = 1.39$ and 1.63 could be decreased. This accomplished if a low-index layer of eighth-wave optical thickness is added to each end of the stack:

$$\text{glass} \quad \frac{L}{2} \quad H\,L\,H\,L\,H\,L\,H \quad \frac{L}{2} \quad \text{air}\,.$$

The spectral transmission of such a stack with $\lambda_o = 700\ m\mu$ is shown in Figure 20.66. The effect of adding the eighth-wave layers is to increase the transmission in the short-wave region. The reflectivity peak at $g = 1.39$ (at 504 $m\mu$ in Figure 20.66) has been decreased from 0.25 to 0.11 and the peak at $g = 1.63$ (430 $m\mu$ in Figure 20.66) is barely perceptible. The reflectivity in the long wavelength region has actually been increased slightly from .25 to .27. The addition of the eighth-wave layers to the seven-layer stack has decreased R_{max} from .95 to .94. Additional layers can be added to this nine-layer stack to increase the transmission even further in the short-wavelength region. Epstein[35] elaborates on methods of accomplishing this. As an example of how the transmission can be increased in the long-wave region, consider the multilayer

$$\text{glass} \quad \frac{H}{2} \quad L\,H\,L\,H\,L\,H\,L \quad \frac{H}{2} \quad \text{air}\,.$$

The spectral transmission of this multilayer is shown in Figure 20.67. The transmission in the red and infrared is quite high but the transmission in the near ultraviolet is considerably lower than that of a quarter-wave stack LHLHLHL. This is usually the case, that when the transmission is increased on the long-wave side of the high-reflectance zone, it is decreased on the short-wave side, and vice versa. In practice, the ultraviolet transmission of the stack shown in Figure 20.67 would be much lower than the computed values if zinc sulfide were used as the high-index layer material, since this material absorbs strongly below 400 $m\mu$. The design of the multilayer shown in Figure 20.67 can also be written as:

$$\text{glass} \quad \left(\frac{H}{2}\ L\ \frac{H}{2}\right)^4 \quad \text{air}\,.$$

This can be regarded as a multilayer with a periodic structure with $\left(\frac{H}{2}\ L\ \frac{H}{2}\right)$ as a basic period. Equation (46) still applies and consequently a high-reflectance zone can occur when an integral number of half-waves fit into the basic period. When the design of the stack can be written in this manner, it can be shown that at any wavelength outside of the high-reflectance zone, it is possible to replace the entire stack of nine layers by a single layer with a ficticious index n_h and ficticious optical thickness τ. That is, for the purposes of computing the reflectivity (outside of the high-reflectance zone), the nine-layer stack is equivalent to a single-layer of index n_h and optical thickness τ. The index n_h is called the Herpin equivalent index. Space does not permit us to describe how the concept of the Herpin equivalent index is used to design multilayer combinations which have a high transmission outside of the high-reflectance zone. For further details, the reader can refer to Weinstein[5] or Epstein[35]. To cite some additional examples, eighth-wave layers of high-index material have been added to the quarter-wave stacks shown in Figures 20.73, 20.78, 20.80, 20.81, and 20.82 to increase the transmission in the long-wave region. The multilayers shown in Figures 20.82 and 20.83 are quarter-wave stacks modified so that the transmission is optimized in the short-wave region. An eighth-wave layer of low index material is added to one end of the quarter-wave stack (next to the incident medium), while a layer of optical thickness $1.28\ \lambda_o/4$ is inserted between the quarter-wave stack and the substrate.

20.4.8.4 The refining method.[36] The refining method consists of varying by a small amount the thickness of each of the layers of a multilayer of specified design, so that the transmission is increased (or decreased, as the case may be) at certain specified wavelengths. This method can be used to increase the transmission of a quarter-wave stack on either the short-wave or long-wave side of the high-reflectance zone. As is mentioned in 20.4.8.1.1, the thickness of each of the layers can be changed by as much as 10% without seriously affecting the reflectivity in the high-reflectance zone. The refining method is similar to the relaxation method which is used to solve complex engineering problems. The computations are sufficiently lengthy and tedious that it is necessary to employ an electronic digital computer. As an example of the application of the refining method, suppose it is desired to increase the transmission of a quarter-wave stack which transmits in the blue but attenuates longer wavelengths. The computed reflectivity of a modified quarter-wave stack is shown in Figure 20.68. The effect of using a large number of layers is to achieve a high attenuation in the green and red $(T_{min} = .001)$ and a sharp cutoff at

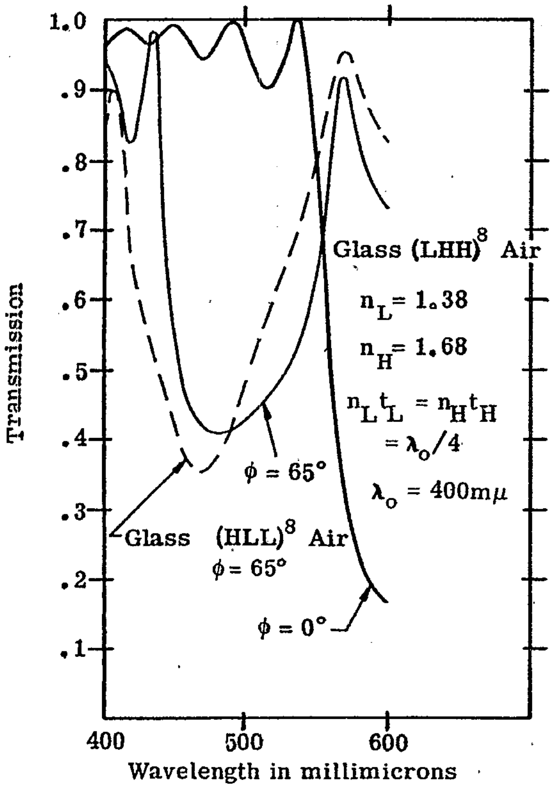


Figure 20.64- T_{av} of a sixteen-layer 2:1 stack at $\phi = 0$ and 65°. The ϕ = 65° curve from Fig. 64 is shown as a dashed line.

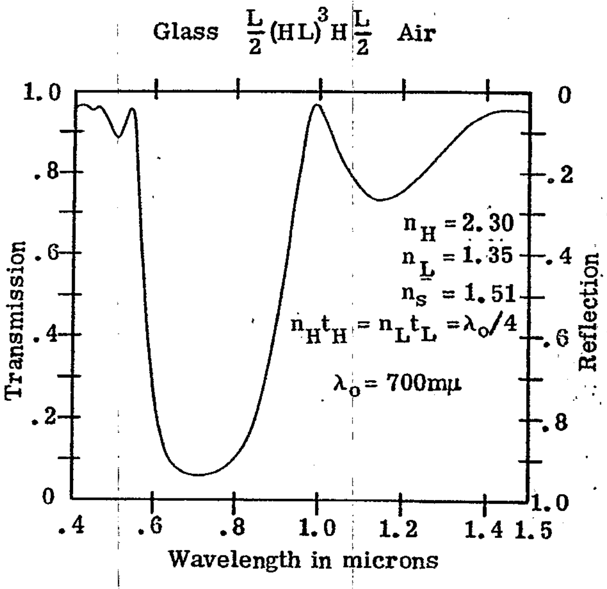


Figure 20.66- Computed spectral transmission of a short-wave pass filter consisting of a modified quarter-wave stack.

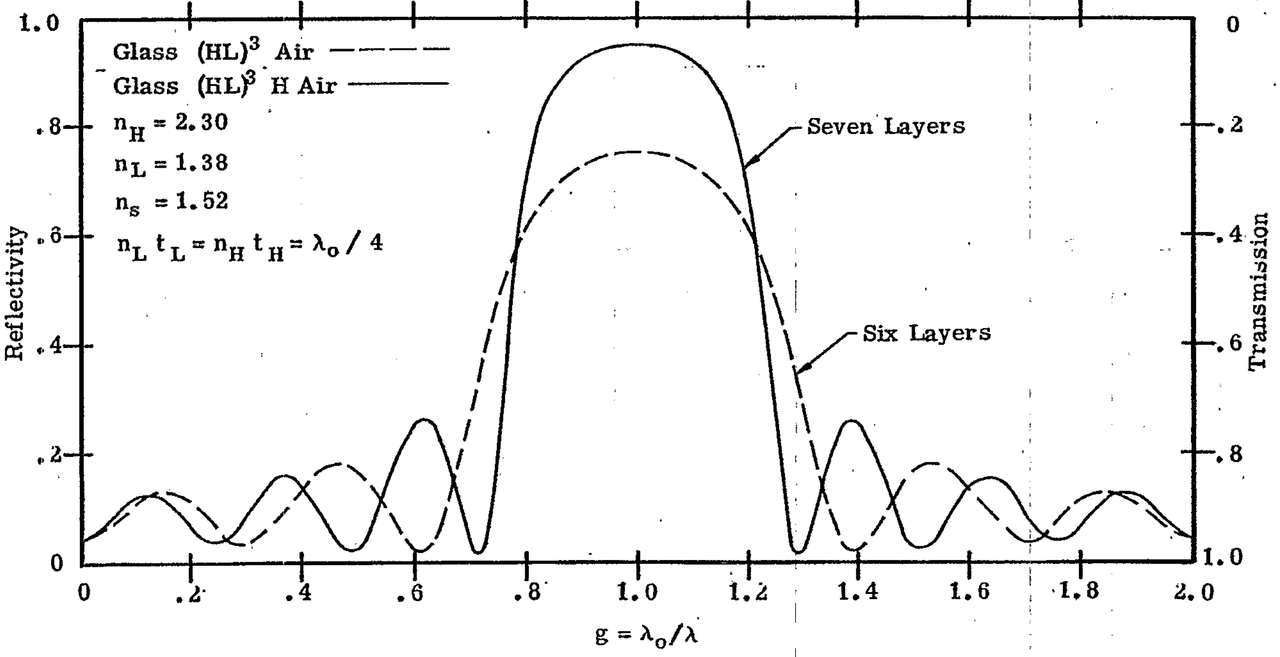


Figure 20.65- Computed spectral reflectivity of six-layer (dashed line) and a seven-layer (solia line) quarter-wave stack.

500 mμ . The reflectivity of the modified quarter-wave stack (DESIGN I) attains a maximum of .33 in this region. This multilayer would be much more effective as a blue pass filter if the undesirable peaks in the reflectivity at 484 mμ and 446 mμ could be eliminated. This is accomplished by varying the thickness of each of the layers by a relaxation process. Table 20.4 shows the quarter-wave optical thickness of each of the layers of the initial stack (Design I) and the refined design (Design II). The maximum change in the thickness of any of the layers is only 12% and the thickness of most of the layers has been changed by only a few percent. This mismatch in the optical thickness of the layers has caused the peak reflectivity to decrease from .9990 (in Design I) to .9988 (in Design IX). Also, the steepness of the transition from the high-reflectance zone to the pass region has been slightly decreased. As another example, the curve designated as Design III in Figure 20.69 depicts the spectral transmission of a quarter-wave stack composed of germanium and silicon monoxide which is used as a long-wave pass filter for the infrared. Table 20.4 shows the thickness of each of the layers (as a fraction $\lambda_0/4$) of the refined multilayer (Design IV) whose transmission curve is also shown in Figure 20.69 .

20.5 LONG-WAVE PASS FILTERS

20.5.1 General properties. From the discussion in 20.4 about the properties of a multilayer with a periodic structure, the method of designing either a long-wave or a short-wave pass filter is fairly obvious. One simply chooses a quarter-wave stack or other type of stack with periodic structure, so that the high-reflectance zone covers the region to be attenuated. It is necessary to choose the materials which are used in the stack and also the number of layers. This can be accomplished after the properties of the multilayer stack have been specified. In establishing the specifications of a long-pass, short-pass, or a band-pass multilayer filter, some of the following properties are considered:

(1) The optical density in the attenuation region. This is discussed in 20.5.1.1.

(2) The transmission in the pass region. In 20.4.8.3 two methods of enhancing transmission in the pass region are presented.

(3) The steepness of the cutoff.* This is discussed in 20.5.1.2.

(4) The change of the transmission with angle, i.e. the angle shift. If a minimum amount of angle shift is required, then high-index materials should be used in the stack, as is mentioned in 20.4.7.

(5) The change of transmission with temperature. In certain applications multilayer filters are used in environments which are either warmer or cooler than room temperature, and the shift of the spectral transmission must be taken into account. Multilayer filters are sometimes placed in thermal contact with an infrared detector which is cooled with liquid nitrogen. Figure 20.79 shows the shift with temperature of the transmission of an infrared filter.

20.5.1.1 The optical density in the attenuation region increases as the index mismatch between the layers of a stack increases and also as the total number of periods increases. In the special case of a quarter-wave stack, the maximum attenuation can be computed from Equations (41) through (43).

20.5.1.2 The sharpness or the steepness of the cutoff * is also an important parameter. The cutoff is the region in which the transmission drops from "high" values in the pass region to "low" values in the attenuation region (in the high reflectance zone). The words low and high in the preceeding sentence are enclosed in quotation marks because the criterion of what constitutes a low or high transmission is rather arbitrary. For example, in the multilayer shown in Figure 20.76 the high value is chosen as 0.70 and the low value as .05. Similarly, the wavelength at which the transmission has decreased to some arbitrary value is called the cutoff wavelength. For example, in Figure 20.79 the wavelength of the $T = .05$ point is chosen as the cutoff wavelength. Regardless of what criterion is chosen, the steepness of the cutoff increases as the number of periods increases. Let us apply as a criterion of the sharpness of the cutoff the wavelength difference $\Delta\lambda$ between the $T = 0.9$ point and $T = 0.1$. Each of the multilayers, whose transmission curves are shown in Figures 20.66 and 20.68, contain the same materials, but have a different number of periods. The $\Delta\lambda$ for the nine-layer stack (Figure 20.66) is 90 mμ, whereas the $\Delta\lambda$ for the seventeen-layer stack (Figure 20.68) is 20 mμ. Although the comparison does a slight unjustice to the nine-layer stack because its cutoff is at longer wavelengths, nevertheless this serves to illustrate the point that the steepness of the cutoff increases at the number of basic periods increases. The order of interfer-

---

* Some writers use the word "cuton" to denote the onset of the attenuation region and "cutoff" to denote the "cutoff" (or stopping) of the attenuation region. In section 20, this distinction is not made.

LAYER	INDEX OF LAYER	QUARTER-WAVE OPTICAL THICKNESS IN mμ	
		DESIGN I	DESIGN II
AIR	1.00	MASSIVE	
1	1.38	300	306
2	2.30	599	639
3	1.38	599	630
4	2.30	599	610
5	1.38	599	606
6	2.30	599	597
7	1.38	599	582
8	2.30	599	573
9	1.38	599	577
10	2.30	599	589
11	1.38	599	596
12	2.30	599	590
13	1.38	599	585
14	2.30	599	601
15	1.38	599	646
16	2.30	599	672
17	1.38	599	616
GLASS	1.52	MASSIVE	

LAYER	INDEX OF LAYER	OPTICAL THICKNESS OF LAYER IN UNITS OF $\lambda_o/4$	
		DESIGN III	DESIGN IV
AIR	1.00	MASSIVE	
1	4.00	0.5	0.419
2	1.80	1.0	1.181
3	4.00	1.0	1.272
4	1.80	1.0	0.967
5	4.00	1.0	0.879
6	1.80	1.0	1.060
7	4.00	1.0	1.194
8	1.80	1.0	1.023
9	4.00	1.0	0.874
10	1.80	1.0	0.961
11	4.00	1.0	1.143
12	1.80	1.0	0.979
13	4.00	0.5	0.405
GLASS	1.52	MASSIVE	

Table 20.4- The design of multilayer filters, whose transmission and reflectivity curves, are shown in Figures 20.68 and 20.69 .

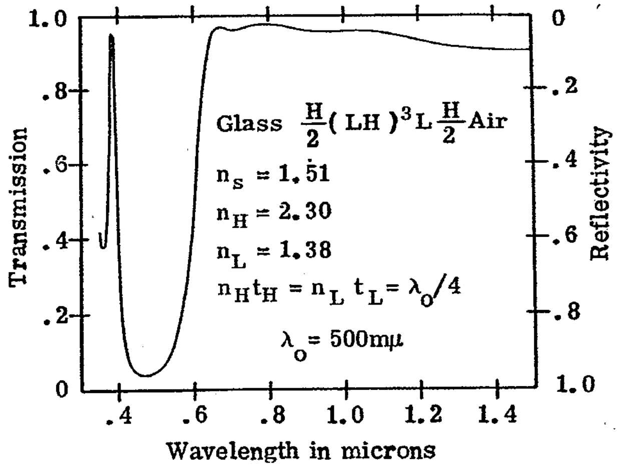


Figure 20.67- Computed spectral transmission of a long-wave pass filter consisting of a modified quarter-wave stack.

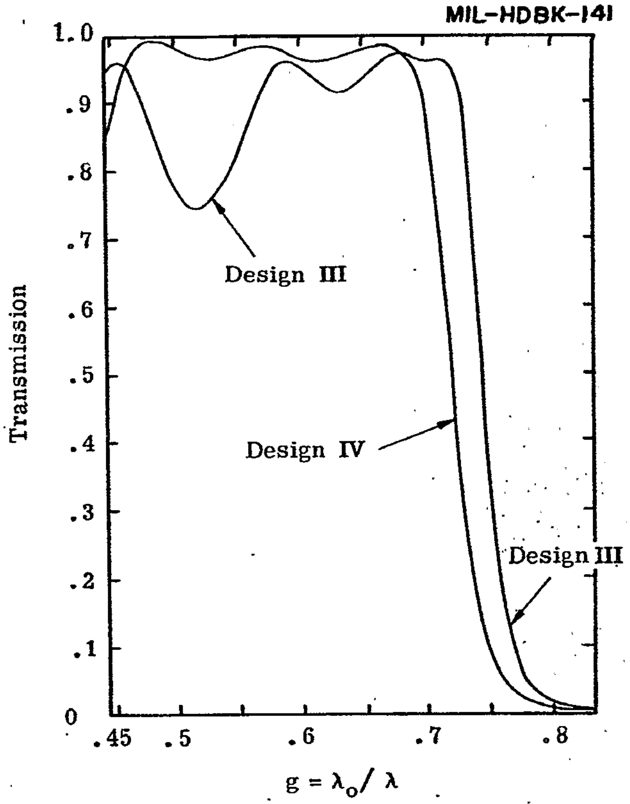


Figure 20.69- Computed spectral transmission of multilayers designated as Design III and Design IV in Table 20.4.

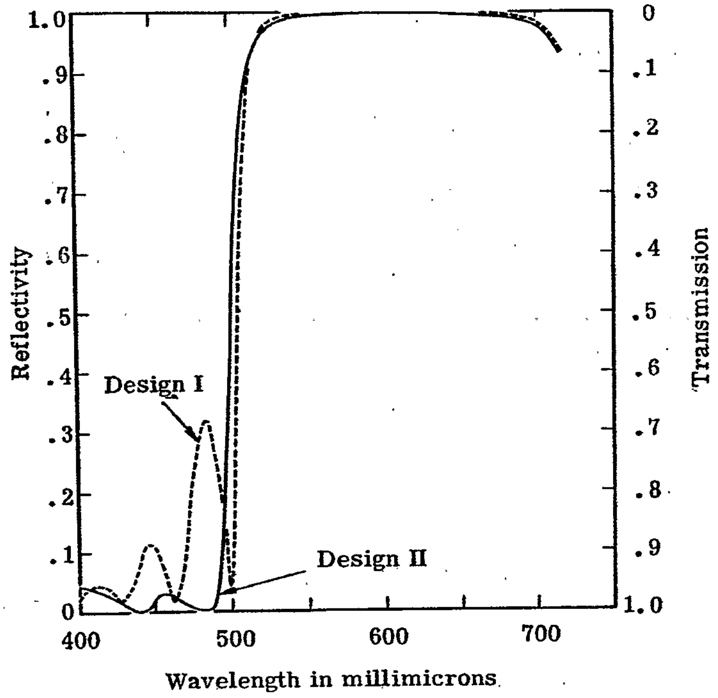


Figure 20.68- Computed spectral reflectivity of the multilayers designated as Design I (dashed line) and Design II (solid line) in Table 20.4.

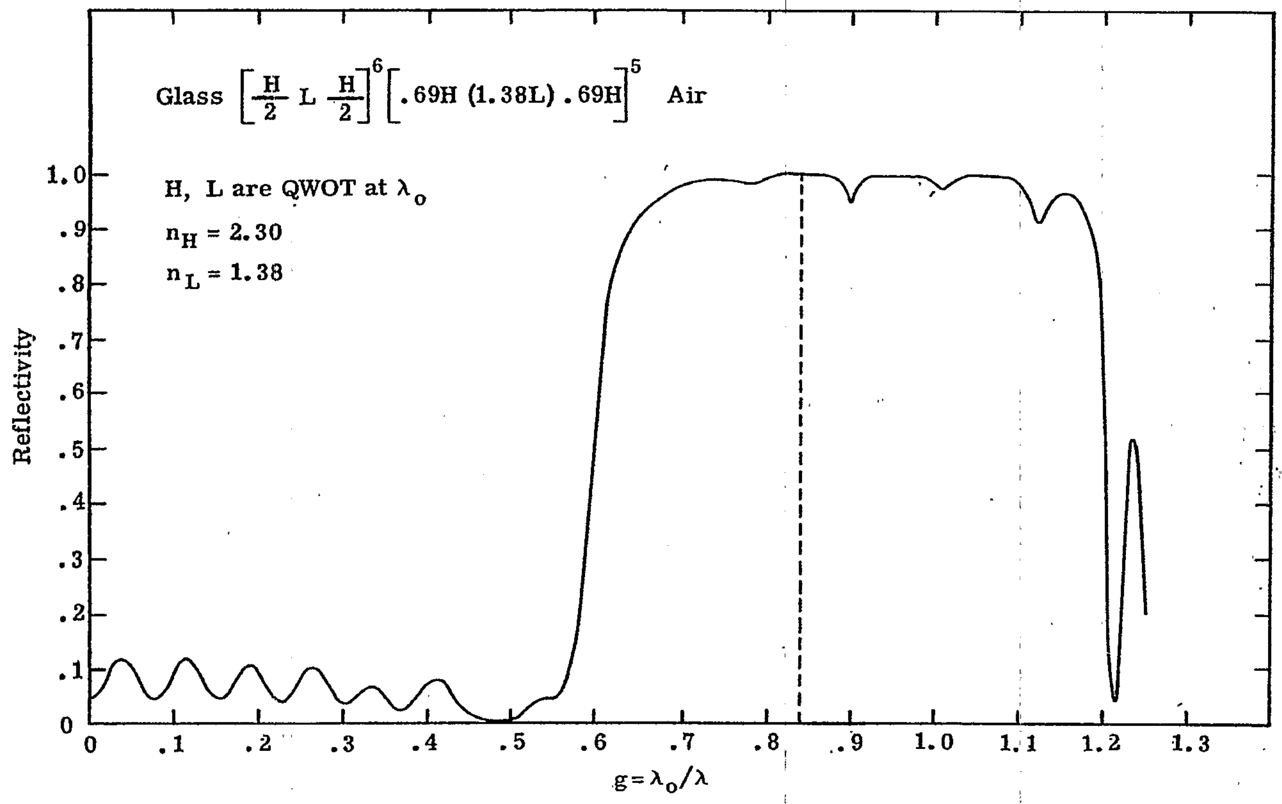


Figure 20.70- Computed spectral reflectivity of a multilayer consisting of an ensemble to two stacks.

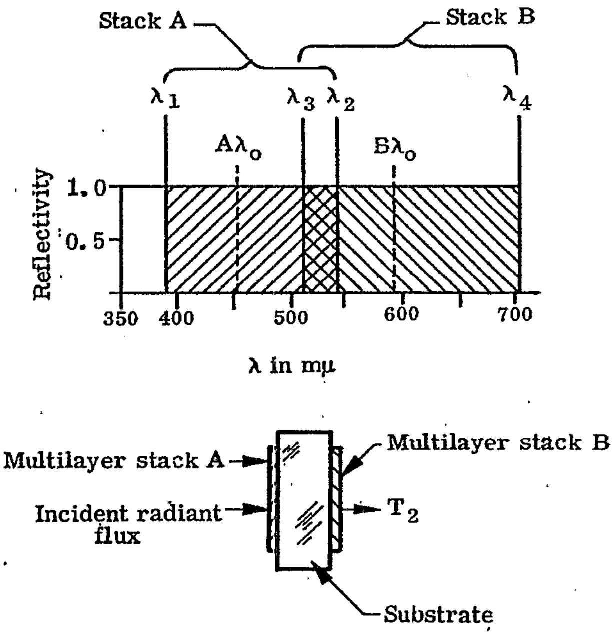
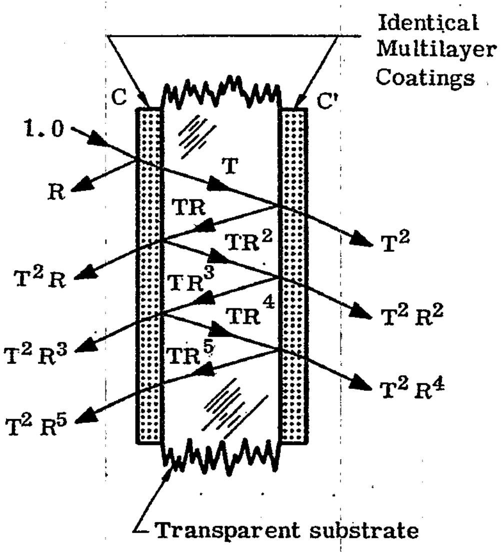


Figure 20.71- The positions of the high-reflectance zones of the two multilayers which constitute a composite filter.



$$T_2 = T^2\left[1 + R^2 + R^4 + \ldots\right]$$

$$= \frac{T^2}{1 - R^2} = \frac{(1 - R)^2}{(1-R)(1+R)}$$

$$= \frac{1 - R}{1 + R} = \frac{T}{1 + R}$$

Figure 20.72- An illustration of the method of computing the total transmission, T_2 of a composite filter consisting of two identical non-absorbing coatings.

ence of the high-reflectance peak influences the steepness of the cutoff. At any given wavelength, a higher-order reflectance peak has a sharper cutoff than a first order peak. The disadvantage of using a higher-order reflectance peak is that the width of the region of high transmission at longer wavelengths is decreased. However, if a first-order high-reflectance peak is used to provide the attenuation region, then there are no regions of high-reflectivity between the long-wave cutoff and infinite wavelength. For example, reflectance peak at 500 mμ of the multilayer shown in Figure 20.67 is first order. In this case the computed transmission of this multilayer is greater than 0.85 mμ in the spectral region from 650 mμ to wavelengths as long as 50 μ in the infrared, or for wavelengths in the microwave region for that matter. Of course in practice the glass substrate and the materials in the multilayer would absorb strongly at wavelengths much shorter than 50 μ.

20.5.1.3 Attenuation over a broad spectral region. It was shown in 20.4.1.4 that the width of the high-reflectance zone is determined by the index ratio of the two materials which constitute the stack and hence the width of the high-reflectance zone is limited by the index of available thin film materials. The optimum width is achieved with a quarter-wave stack which has a large index mismatch. For example, in the region from 300 to 400 mμ the non-absorbing thin film materials have an index which does not exceed 2.10. The optimum width of the high-reflectance zone is achieved if this high-index material is used in conjunction with a low index material such as cryolite (index 1.35). Sometimes it is desired to attenuate a spectral region which is greater than this optimum width. One obvious solution is to combine two or more such stacks. The thickness of the layers in each of the stacks is chosen so that the high-reflectance zone of each one covers a different part of the spectral region to be attenuated. Here the term stack has been used to denote a group of films which have periodic structure. There are two ways of arranging the stacks so that a broad attenuation region is achieved.

(1) The stacks are deposited upon the same substrate, that is, they are piled on top of each other. Since this type of multilayer is an ensemble of individual stacks, it will be referred to as an ensemble multilayer.

(2) Each stack is deposited on a different surface. This configuration will be referred to as a composite filter.

20.5.1.3.1 Ensemble of stacks. As an example of a multilayer which is an ensemble of two individual (modified) quarter-wave stacks, consider.

$$\text{glass}\ \left(\frac{L}{2}\ H\ \frac{L}{2}\right)^6\ \left(\frac{L'}{2}\ H'\ \frac{L'}{2}\right)^5\ \text{air}$$

where the index of the L land H layers is 1.38 and 2.30, respectively. Using our terminology, the group of layers $(L/2\ H\ L/2)^6$ constitutes one stack and the group $(L'/2\ H'\ L'/2)^5$, another stack. If the optical thickness of the L, H and L', H' layers is chosen in the ratio of 1.38: 1, then the high-reflectance zones to the two stacks are contiguous. The reflectivity versus $g = \lambda_0/\lambda$ of this combination is depicted in Figure 20.70. The dashed line shows the boundary of the high-reflectance zones of each stack. The attenuation region covers nearly an octave. In general, two difficulties are encountered in constructing multilayers of this type. First, interference effects often occur between the various stacks and the ensemble filter acts like a Fabry-Perot type filter (described in 20.10), to the extent that narrow transmission bands appear in the attenuation region. Second, such filters are often difficult to manufacture. Each of the films in the filter has a small amount of mechanical stress (refer to 20.2.4.2.4) and if the number of layers is large, the total stress can build up to the point where the multilayer will no longer adhere to the substrate. Multilayers of this type which have a high-reflectivity over a wide spectral region are often called broad-band reflectors. The spectral reflectivity of such a reflector is shown as curve B in Figure 20.95.

20.5.1.3.2 The composite filter. The composite filter consists of two or more individual multilayers which are deposited on separate substrates and arranged so that the incident radiant flux passes through each one. Because the multilayers are on separate substrates, it is possible to separate them physically by a distance of many thousands of wavelengths, thus avoiding interference effects between the various multilayers. If only two multilayers are used in a composite filter, it is often convenient to deposit each multilayer on a side of the substrate, as is shown in Figure 20.71. As an example of this procedure, suppose a filter is required which attenuates the entire visible spectral region from 400 mμ to 700 mμ. Since a single quarter-wave stack of zinc sulfide and magnesium fluoride has a high-reflectance zone which covers approximately one-half of this region, a composite filter composed to two quarter-wave stacks is requisite. Each stack is deposited on a side of the glass substrate, as is shown in Figure 20.71. The high-reflectance zone of multilayer stack "A" extends from λ_1 to λ_2 and the zone of stack "B" from λ_3 to λ_4. Figure 20.71 shows how the high-reflectance zones are arranged so they not only cover the spectral region to be attenuated but also overlap in the center so that there is no possibility of a transmission leak. Stack "A" has a QWOT = 453 mμ and stack "B" a QWOT = 594 mμ. These values are chosen so that region of attenuation extends from 390 mμ to 706 mμ. Having determined the QWOT and materials of the two stacks, the only remaining problem is to choose the number of periods. This number is determined by the amount of attenuation required, the steepness of the cutoff required at 700 mμ, or by economic considerations. The latter point should be considered if the filter is to produce in large numbers; the cost of manufacture depends upon the number of layers in the stack. It is noted that the transmission beyond 700 mμ would be enhanced if eighth-wave layers were added to each stack, so that stacks "A" and "B" are similar to the multilayer shown in Figure 20.67. The addition of these eighth-wave layers shifts the wavelength of maximum reflectivity of each stack, but does not affect the position of the high-reflectance zones.

20.5.1.4 Since the multilayers which are composed of dielectric materials have a negligible amount of absorption, the transmission properties of these filters, when placed in tandem, are considerably different from absorption-type filters. As an example, suppose that an absorption-type filter consisting of a slab of colored glass has a transmission T of 0.01 at some wavelength, say 610 mμ. If an identical filter is placed in tandem, then the transmission T_2 of the entire system is reduced to the low value of

$$T_2 = T^2 = 0.0001 .$$

Next, consider the case of a multilayer filter which is composed of dielectric (non-absorbing) layers, so that $T = 1 - R$. For purposes of comparison, suppose that this multilayer is designed so that it also has a $T = .01$ at 610 mμ. If an identical multilayer is deposited on the opposite side of the same transparent substrate, as is shown in Figure20.72, then the total transmission T_2 of the tandem arrangement is close to 0.005, or more precisely:

$$\begin{aligned} T &= (1 - R)/(1 + R) \\ &\simeq T/2 \text{ as } R \rightarrow 1.0 \end{aligned} \tag{48}$$

The reason for this behavior can be seen in Figure 20.72, which shows identical multilayer coatings C and C' deposited on each side of the transparent substrate. Since the substrate is quite thick, interference effects can be neglected and the intensity of the beams which emerge can be added. As is shown in Figure 20.72, a radiant flux proportional to $T = 1 - R$ penetrates into the medium (i.e. the substrate) between the two multilayers. If the reflectivity is close to 1.0, then this flux is trapped and bounces back and forth between the two multilayers many times, decreasing in intensity only a small amount at each reflection. Eventually all of the flux escapes through interfaces C and C'. If R is close to 1.0, then approximately one half of it escapes through C and one-half through C'. Since T_2 is proportional to the radiant flux which penetrates C', T_2 is decreased by only a factor of two. The total transmission T_2 is decreased below the value given in Equation (48) if there is a small amount of scattering in the multilayers or if there is absorption in either the multilayer or the substrate. T_2 can also be decreased by using a wedge-shaped substrates so that the rays "walk off" the ends of the filter.

20.5.2 Long-wave pass color filters. From the discussion in 20.5.1 and 20.4.8.2.1, it is evident that it is a straightforward task to produce filters for the visible spectral region which pass in the long-wave region. However, it should be remembered that in the visible region there are many absorption filters of colored glass or organic dyes which rival multilayers in the sharpness of their cutoff, high attenuation in the short-wave region and high transmission at long wavelengths. Not only are these absorption filters often less expensive than a multilayer filter, but they have the additional advantage that they are virtually free from the variation of the transmission with the angle of incidence (i.e. the angle shift) which is inherent in multilayer filters. However, multilayer filters have the advantage that they can be manufactured to any specification so that the cutoff can be positioned at any wavelength, whereas the cutoff of glass and dye absorption filters is only at wavelengths which nature has provided. As an example of a long-wave pass color filter, Figure 20.73 shows the measured spectral transmittance of a modified quarter-wave stack composed of fifteen layers of zinc sulfide and magnesium fluoride. Eighth-wave layers of zinc sulfide on either end of the stack increase the transmission in the long-wave region. The transmittance of this filter is similar to the computed curve shown in Figure 20.67.

20.5.3 The cold mirror. The cold mirror is a multilayer composed of dielectric materials which has a high-reflectivity in the visible spectral region. The reflectivity drops abruptly at 700 mμ so that a high transmission is achieved in the near infrared. The discussion in 20.1.2.7 describes briefly how this cold mirror can be placed behind a light source in a projection system so that the light in the visible spectral region is deflected towards the lens, but the heat, (the radiant energy in the infrared), passes out of the system. There are many designs for cold mirrors; they usually consist of an ensemble of stacks, as described in 20.5.1.3.1. An effective cold mirror has a smooth reflectivity curve in the visible region so that the color of the reflected light is not altered. The sharpness of the cutoff at 700 mμ and the amount of transmission in the infrared are also important. A cold mirror which is used in conjunction with an arc lamp should have a sharp cutoff, since a typical arc lamp has a larger portion of its radiant energy concentrated in the region between 700 mμ and 800 mμ than does a tungsten filament. Since 85% of the radiant energy of a tungsten incandescent lamp operated at 3350°K is outside of the visible region, a combination of a cold mirror and a heat reflector could theoretically reduce the heat six-fold. In practice, a two-fold to three-fold reduction is achieved by using a cold mirror and heat reflector in place of an aluminum mirror. Figures 20.74 and 20.75 depict the spectral transmittance and reflectance of some cold mirrors which are manufactured commercially. Both Dimmick[37] and Turner[38, 39a] describe the use of cold mirrors in projection and illumination systems. A basic U.S. patent on the cold mirror has been issued to Koch[39b].

20.5.4 Long-wave pass filters for the infrared. Multilayer filters are used extensively in the infrared as components of missiles which have a heat-seeking guidance mechanism, infrared surveillance systems, and in the spectrochemical analysis of organic vapors. In the infrared, a limited number of absorption-type filters are available and hence multilayer filters are widely used. Infrared multilayers can be composed of the same materials which are used in the visible, such as zinc sulfide and chiolite. However, the use of certain materials is restricted by the stress in the films (see 20.2.3.2.4). Films of germanium can be used at wavelengths longer than 1.3 μ and lead telluride beyond 3.9 μ. Both of these materials have high refractive index. As is illustrated in Figure 20.53, a stack which has a quite broad high-reflectance zone is obtained when either of these materials is used in conjunction with a material of low refractive index, such as chiolite or silicon monoxide. The spectral width of the attenuation region can be increased by using two or more stacks, combined either as a ensemble (20.5.1.3.1) or as a composite filter (see 20.5.1.3.2). Figures 20.76 to 20.81 show the spectral transmittance of some long-wave pass filters which are manufactured by several commercial firms. It is interesting that the cutoff region of the stack shown in Figure 20.79 is provided by a 3:1 stack, which has the advantage of having a lower angle shift than the quarter-wave stack (see Section 20.4.7). It might be conjectured that the spike in the transmission at 3.8μ of the multilayer shown in Figure 20.80 is due to the fact that the edge high-reflectance zone occurs at this wavelength but that the transmission remains low because the lead telluride films are absorbing in this region. Using the refractive indices in Table II, from Equation (39) it is found that the ratio $\lambda_2/\lambda_1 = 1.66$, where λ_2 and λ_1 are the wavelengths at the long and short wave edge of the high-reflectance zone. From Figure 20.82 we see that $\lambda_2 = 6.5\ \mu$, and thus $\lambda_1 = 3.9\ \mu$; this conjecture is probably correct. The transmission peaks at shorter wavelengths in Figures 20.80 and 20.81 could be eliminated by depositing another long-wave pass multilayer on the opposite side of the substrate, thus forming a composite filter. The transmission in the long-wave region in both of these filters could be improved if an antireflection coating were deposited on the opposite side of the arsenic trisulfide glass substrate.

20.6 SHORT-WAVE PASS FILTERS

In Section 20.5.1, we enumerated some of the properties of long-wave pass filters, such as the attenuation, sharpness of the cutoff, angle shift, and so on. These general considerations apply equally well to short-wave pass filters, with the following important exception: The spectral width of the high-transmission region of a quarter-wave stack or other stack with a periodic structure is limited by the fact that at shorter wavelengths higher order high-reflectance zones always occur. For example, suppose that the 3:1 stack shown in Figure 20.59 is used to attenuate in the 1.8 μ to 2.2 μ region. Additional periods can be added to the stack if a higher attenuation in this region is required. The pass region on the short-wave side extends from 1.15 to 1.7 μ. At 1.15 μ the second-order high-reflectance peak occurs and the multilayer attenuates from 0.9 to 1.15 μ. It is evident from Figures 20.56, 20.57, and 20.58 that these higher order attenuation regions are not confined to the 3:1 stack, but occur in all stacks which have a periodic structure, regardless of the composition of the basic period. A comparison of Figures 20.56, 20.57, and 20.58 shows that the quarter-wave stack has the widest region of high transmission on the short-wave (high-frequency) side of the first order high-reflectance zone. The third-order high-reflectance zone occurs at one-third of the wavelength of the first order zone. The spectral width of high transmission of the 2:1 stack is somewhat smaller; the second-order high-reflectance zone occurs at one-half the wavelength of the first order zone. However, we have considered only the case where two layers are used in the basic period of the stack. If more than two layers are used in the basic period of a stack with a periodic structure, then a short-wave pass spectral region which is even wider than that of a quarter-wave stack can be achieved.[40] Regardless of whether a quarter-wave, 2:1 or 3:1 stack is used, it is desirable to add some additional layers to the stack to enhance the transmission in the short-wave region, as is shown in Figure 20.66.

20.6.1 Short-wave pass color filters. Multilayer filters are particularly useful as short-wave pass filters in the visible spectral region, because most of the colored glass and dyed gelatin filters have a low transmission in the pass region and a cutoff which is not sharp. For example, a typical blue glass absorption filter which transmits below 480 mμ has a peak transmission of 0.38 at 410 mμ and a bell-shaped transmission curve. Also, most absorption filters which transmit the blue and near ultraviolet have a leak-a transmission band in the red. Figures 20.66, 20.68, and 20.82 show the spectral transmission of some short-wave pass filters. The multilayer shown in Figure 20.82 is a quarter-wave stack with films added to enhance the transmission in the short-wave region, as described in 20.4.8.3.1.

LONG-WAVE PASS FIFTEEN-LAYER CUTOFF FILTER WITH T=0.50 AT 0.570μ

$G\left[(.5H)L(.5H)\right]^7 A$

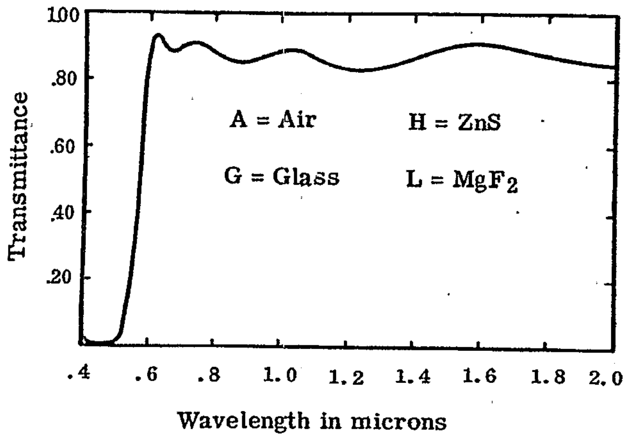


Figure 20.73- Measured spectral transmittance of a long-wave pass filter. Courtesy of Bausch and Lomb, Inc.

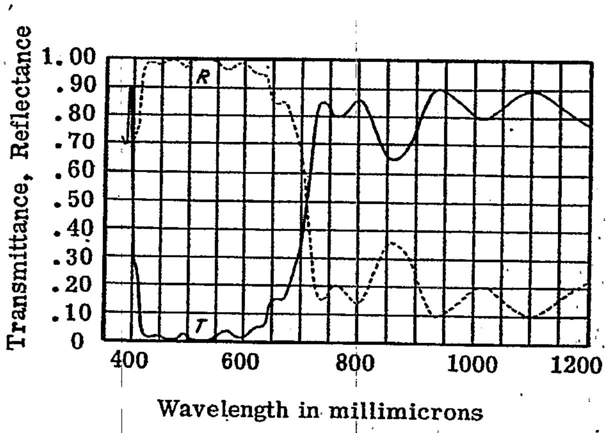


Figure 20.74- Measured reflectance and transmittance of a cold mirror multilayer coating. Courtesy of Balzers Aktiengesellschaft, Liechtenstein

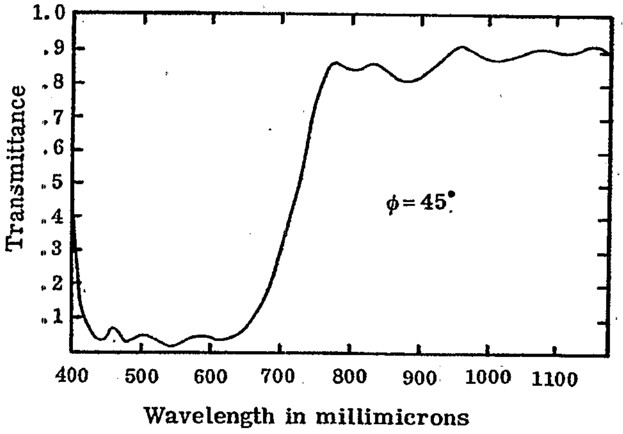


Figure 20.75- Measured spectral transmittance at φ = 45° of a cold mirror of unspecified design. Courtesy of Fish-Schurman, Corporation.

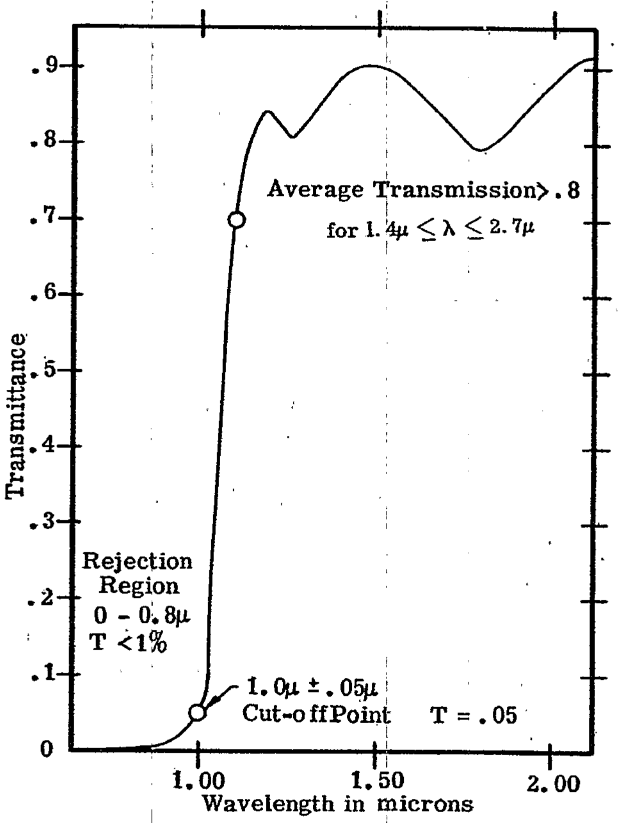


Figure 20.76- Measured spectral transmittance of a long-wave pass filter of unspecified design. Courtesy of Eastman Kodak Company.

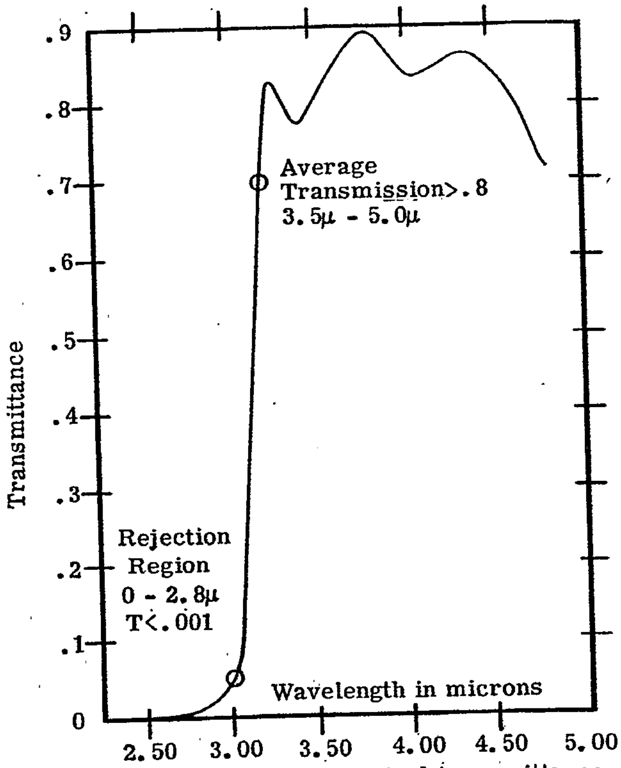


Figure 20.77- Measured spectral transmittance of a long-wave pass filter of unspecified design. Courtesy of Eastman Kodak Company.

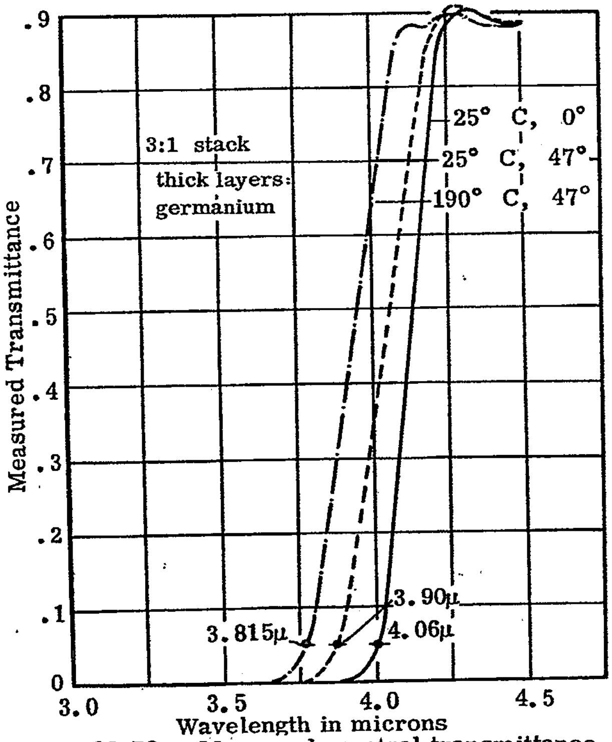


Figure 20.79- Measured spectral transmittance of a long-wave pass filter at $\phi = 0$ and $\phi = 45°$ and at low temperature. The design is a modified 3:1 stack. Courtesy of Optical Coating Laboratory, Inc.

MEASURED SPECTROPHOTOMETRIC TRANSMITTANCE OF A LONG-WAVE PASS ELEVEN-LAYER CUTOFF FILTER

$G\ [(.5H)L(.5H)]^5\ A$

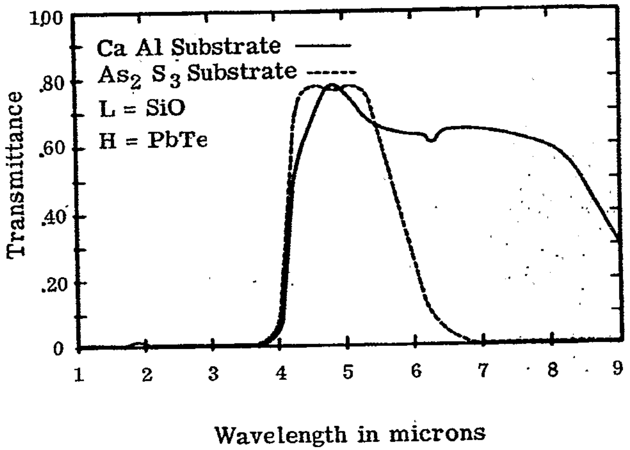
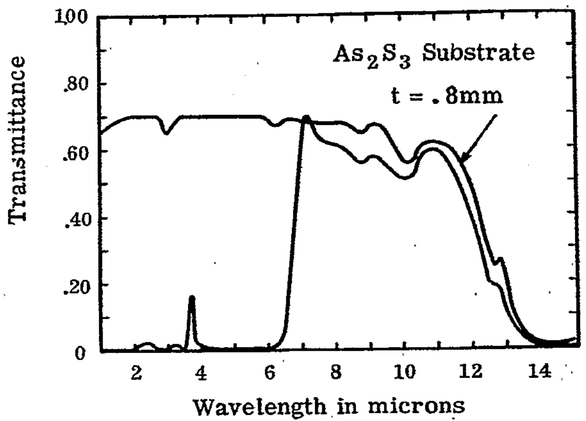


Figure 20.78- Measured spectral transmittance of a modified quarter-wave stack on two different types of substrate, namely As_2S_3 glass (solid curve) and calcium aluminate (dotted curve). Courtesy of Bausch and Lomb, Inc.

LONG-WAVE PASS NINE-LAYER CUTOFF FILTER AT 6.5 MICRONS

$G\ [(.5H)L(.5H)]^4\ A$



A = Air
L = ZnS
G = As_2S_3 Glass
H = PbTe

Figure 20.80- Measured spectral transmittance of a long-wave pass filter on an As_2S_3 glass substrate. The transmission of both the filter and the bare substrate would be improved if the substrate were antireflection coated. Courtesy of Bausch and Lomb, Inc.

20.6.2 Heat reflectors. Section 20.1.2.7 describes briefly how a multilayer heat reflector deflects the heat from a projection lamp away from the film gate. For many years heat absorbing glass has been used to absorb the radiant energy in the infrared, but this has the disadvantage that the heat must be removed from the glass by air cooling or other means. If this is not done, the glass becomes quite hot and often fractures. Multilayer heat reflectors have the advantage that they reflect, rather than absorb, the near infrared and thus remain comparatively cool. Figures 20.83, 20.84, 20.85, 20.86, and 20.87 show the spectral transmission of some infrared reflectors. The multilayer in Figure 20.83 is identical to that of Figure 20.82, with the exception that the QWOT of the former has been moved to longer wavelengths. The multilayer shown in Figure 20.84 attenuates far into the infrared, but unfortunately it also attenuates in part of the red spectral region. The width of the attenuation region in the infrared of the multilayer shown in Figure 20.85 is typical of what can be attained with a simple quarter-wave stack. The multilayer shown in Figure 20.86 not only reflects the infrared, but also has a higher order high-reflectance zone in the blue. Although the design of the multilayer is unspecified, the fact that the maximum of the red reflection band occurs at twice the wavelength of the blue reflection band leads one to suspect that this is a 2:1 stack. It should be pointed out that the heat reflector does not have to be used in conjunction with a cold mirror, as shown in Figure 20.82 Some advantage is gained if only a heat reflector is inserted into the beam at an angle.

20.6.2.1 Another type of heat reflector is used in conjunction with the solar energy cells which are used in satellites and space vehicles. The problem is that only the radiant energy in the range from 0.4 to 1.2 μ produces appreciable electrical energy. The radiant energy outside of this region merely heats the cell, thereby decreasing its efficiency. Therefore, solar cells used in space vehicles are usually protected with a short-wave pass filter which has a high-reflectivity from 1.2 μ out to longer wavelengths. The advantages of using these filters are discussed by Thelen.[41] It is remarked that heat reflecting mirrors are often called hot mirrors. It should be remembered that both the cold mirror (described in Section 20.5.3) and the hot mirror are made of non-absorbing materials. Thus neither of them absorbs an appreciable amount of radiant energy and thus they both remain comparatively cool.

20.6.3 Short-wave pass filters for the infrared. One difficulty which is encountered in using a multilayer as a short-wave pass filter for the infrared is that the spectral width of the pass region at short wavelengths is narrow, especially if high-index materials such as silicon, germanium, or lead telluride are used in the filter. In the latter case, the attenuation region is quite wide, and this width is at the expense of the width of the pass region. One way to avoid this difficulty is to use materials in the stack which have smaller ratio of refractive index, n_a / n_b . Another approach is to use a stack which has a basic period which contains more than two refractive indices.[40]

20.7 BEAM SPLITTERS

20.7.1 A beam splitter is used to divide a wavefront into two portions and direct each portion in a different direction. Beam splitters can be arbitrarily divided into two classifications:

(1) Achromatic, or neutral, beam splitters (see Section 20.7.2).

(2) Color selective beam splitters. These are sometimes called dichroic mirrors and are discussed in Section 20.7.3.

20.7.1.1 The properties of a beam splitter are usually specified by the average transmission, T_{av}, and average reflectivity, R_{av}, (defined in 20.1.3.7). However, as is shown in 20.7.1.2, it is often important to know T_s, T_p and R_s, R_p in the two planes of polarization so that the amount of polarization produced by the beam splitter can be determined. If the beam splitter is used as a component of an interferometer, such as a Michelson or Twyman-Green, then the variation with wavelength of the phase shift upon reflection should be considered, since the position of the fringes in the interferometer depends upon this phase shift. The phase shift of a silver film is shown in Figure 21.17 .

20.7.1.2 Polarization effects. If T_s and T_p are unequal, or if $R_s \neq R_p$, then the beam splitter has a polarizing effect, which can be quite important. For example, a beam splitter which is used in a Michelson interferometer should divide the beam equally in order to produce fringes with maximum intensity. Thus a beam splitter composed of dielectric films which has an R_{av} of .50 might look attractive as a beam splitter for this purpose. However, it is important to know the R and T in each plane of polarization, since the beams interfere separately in each plane of polarization. For example, if a beam splitter for a Michelson interferometer has $R_{av} = 0.50$, but $R_s = 0.10$ and $R_p = 0.90$, then the fringe intensity, which is proportional to the product $R \cdot T$, is quite low. The dissimilarity of the reflectivity in the two planes of polarization can also affect other devices. For example, suppose a beam splitter in a glass cube is mounted behind the objective lens of a camera, as is shown in Figure 20.88. Such a beam splitter is manufactured by evaporating the multilayer coating onto the hypotenuse face of a 45-90-45 glass prism, and then cementing an identical prism onto it. The purpose of the beam splitter is to reflect 20% of the light into

MEASURED TRANSMITTANCE OF AN ELEVEN LAYER LONG-WAVE PASS FILTER ON As_2S_3 GLASS SUBSTRATE

$G\ [(.5H)L(.5H)]^5\ A$

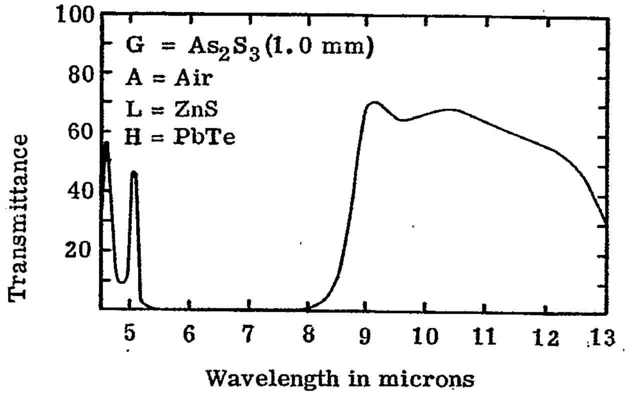


Figure 20.81- Measured spectral transmittance of a long-wave pass filter deposited on an As_2S_3 glass substrate. The transmission beyond 8 μ would improve if the "back" side of the substrate were antireflection coated. Courtesy of Bausch and Lomb, Inc.

SHORT-WAVE PASS THIRTEEN-LAYER CUTOFF FILTER WITH T=.5 AT .572μ

$G\ (.78L)\ [(.5L)H(.5L)]^6\ A$

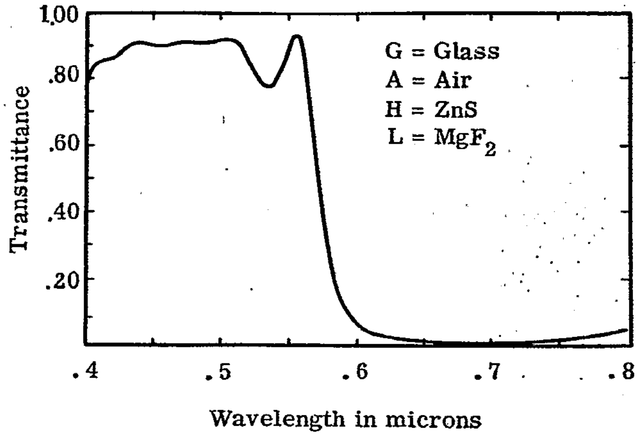


Figure 20.82- Measured spectral transmittance of a short-wave pass filter. Courtesy of Bausch and Lomb, Inc.

SHORT-WAVE PASS THIRTEEN-LAYER CUTOFF FILTER WITH T=.5 AT .80μ

$G\ (.78L)\ [(.5L)H(.5L)]^6\ A$

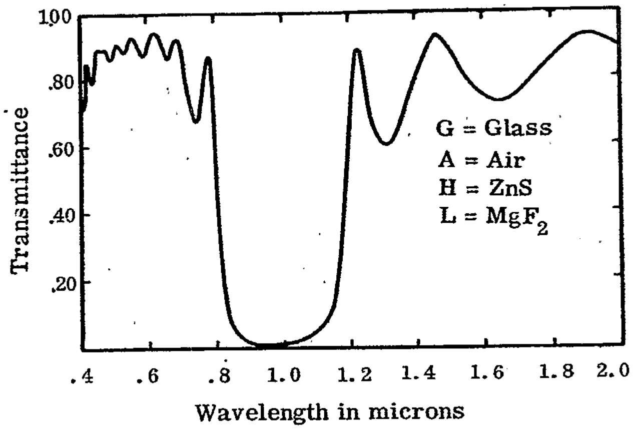


Figure 20.83- Measured spectral transmittance of a short-wave pass heat reflecting filter. Courtesy of Bausch and Lomb, Inc.

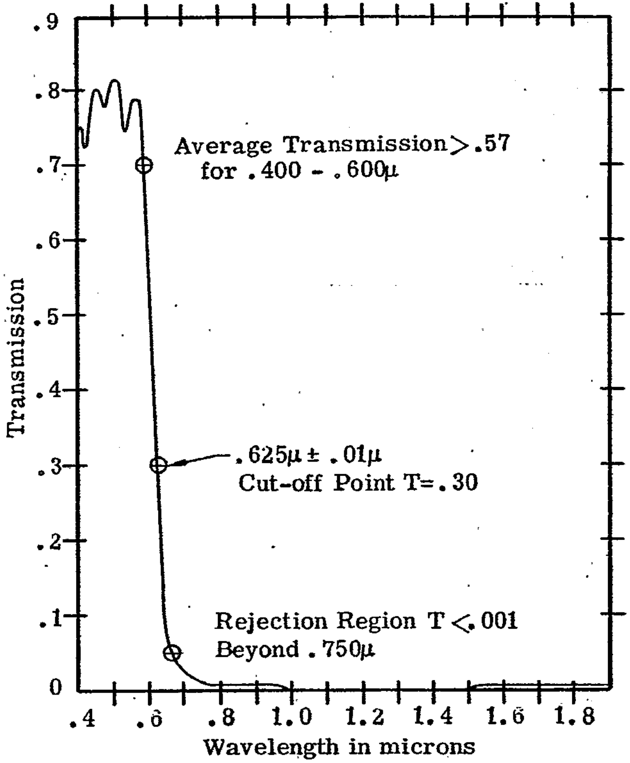


Figure 20.84- Measured spectral transmittance of a short-wave pass filter. Courtesy of Eastman Kodak Co.

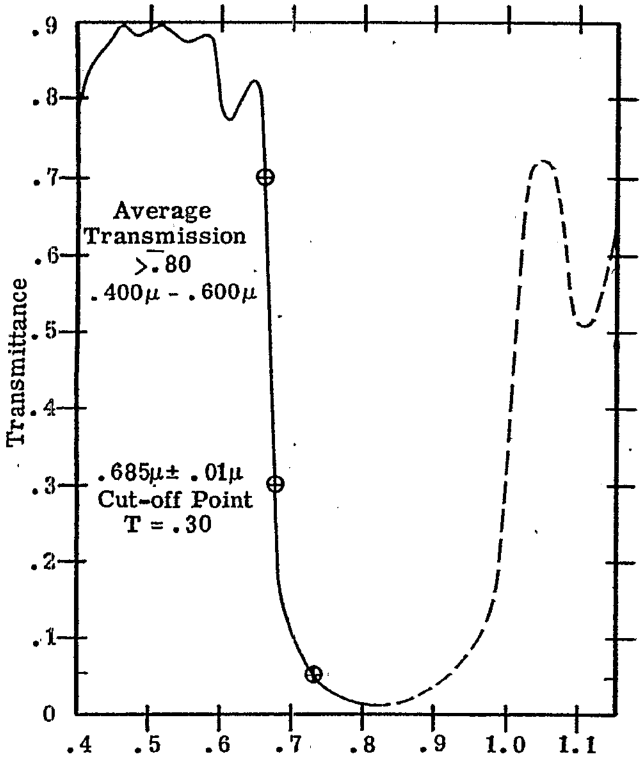


Figure 20.85- Measured spectral transmittance of a short-wave pass heat reflecting filter. Courtesy of Eastman Kodak Company.

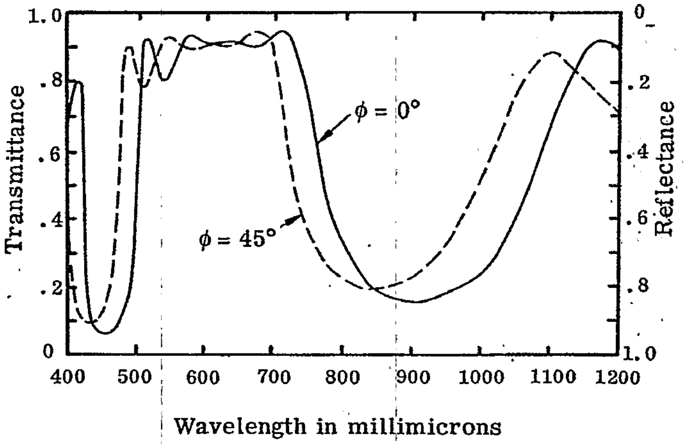


Figure 20.86- Measured spectral transmittance multilayer which reflects the blue and near infrared. Courtesy of Fish-Schurman, Corporation.

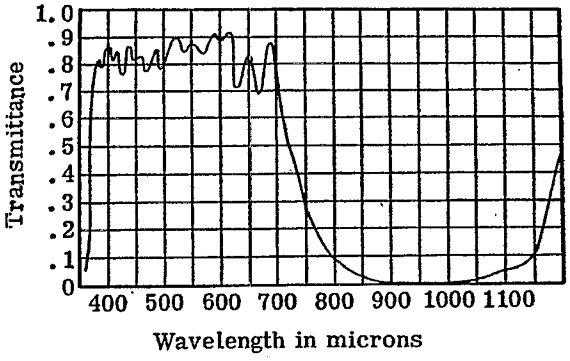


Figure 20.87- Measured spectral transmittance of a multilayer heat reflecting filter. Courtesy of Balzers Aktiengesellschaft, Liechtenstein.

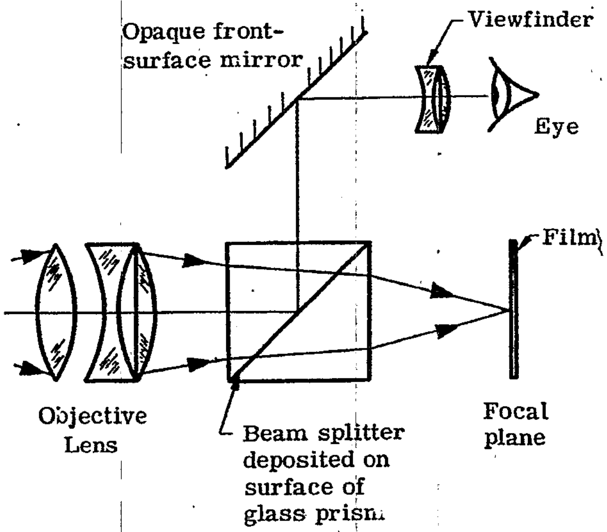

Opaque front-surface mirror
Viewfinder
Eye
Film
Objective Lens
Beam splitter deposited on surface of glass prism
Focal plane

Figure 20.88- The utilization of a beam splitter in a camera to deflect an image into the view finder.

the view finder. As long as the light from the scene which is being photographed is unpolarized, then any beam splitter which has the specified $T_{av} = 0.80$ and $R_{av} = 0.20$ is satisfactory. However, suppose that $R_s = 0.08$ and $R_p = 0.32$ so that the beam splitter is polarizing. If a scene is photographed which produces polarized light, such as certain portions of the sky or the light which is reflected from a wet surface, then the brightness of the scene which is seen in the view finder and the scene which is recorded on the film are quite different. Another effect is that the brightness in the view finder changes as the camera is rotated about the axis of symmetry of the objective lens.

20.7.2 Achromatic beam splitters. An achromatic, or neutral, beam splitter should reflect and transmit equally at all wavelengths. In practice, the transmission of most beam splitters changes slightly with wavelength.

20.7.2.1 Silver films. Thin films of silver are widely used as beam splitters. For many years only chemical methods were used for depositing the films, but more recently superior quality silver films have been produced by evaporation in a vacuum. The optical constants, n and k, vary considerably with wavelength and consequently the R and T of a film of a given thickness also change. The R and T of silver films of varying thickness at a wavelength of 550 mμ is shown in Figure 21.16. Due to the dispersion of the optical constants, silver films which have a given value of t/λ have a much lower absorption in the red than in the blue. Silver films have the advantage that they are easy to prepare. The disadvantages are that they are less efficient than a dielectric beam splitter because they absorb part of the light and that they deteriorate after they are removed from the evaporator. A typical rule of thumb is that silver beam splitter which is designed to divide the light equally reflects 1/3, transmits 1/3 and absorbs 1/3. Sennett and Scott[42] have measured the transmission and reflectivity of some silver films. In the visible spectral region metals other than silver have a high absorption and are generally not used as beam splitters.

20.7.2.2 Dielectric films.

The simplest type of dielectric beam splitter is a single film of zinc sulfide or titanium dioxide deposited on a glass substrate. These beam splitters are neutral in color and divide the light in a ratio of $R : T = 0.4 : 0.6$. Figure 20.89 shows the spectral reflectivity and transmission at $\phi = 45°$ of a single film of TiO_2 on a glass substrate. The dispersion of the refractive index of the TiO_2 has been taken into account in this calculation. The curves in Figure 20.89 give some idea of the flatness of the reflectivity curve and also the amount of polarization which is produced by the beam splitter. Holland[43] describes in detail how beam splitters of this type are prepared. Figure 20.90 shows the spectral reflectivity and transmission at $\phi = 45°$ of two types of beam splitters which are produced commercially. It should be remembered that the uncoated side of the substrate of a beam splitter introduces a transmission loss and also some polarization into the beam. This effect is not large and is reduced even more if the opposite side of the substrate is covered with an antireflection coating.

20.7.3 Color selective beam splitters (dichroic mirrors). Color selective beam splitters are sometimes called dichroic mirrors. One of the definitions of the word dichroism refers to the selective absorption and transmission of light as a function of wavelength regardless of the plane of vibration.

20.7.3.1 It could be argued that section on dichroic mirrors logically belongs in Sections 20.5 and 20.6, which covered long-wave and short-wave pass filters. Actually, this is true, because any of the multilayers described in those chapters could be used a dichroic mirror simply by tilting them at an angle. The difference is that when these multilayers were used as pass filters, we were concerned only with the transmitted light and gave little attention to what happened to the reflected light. If the multilayer is used as a beam splitter, then the reflected light is utilized. Color selective beam splitters are used extensively in optical systems where different kinds of information are identified by a different color and combined so that they are projected simultaneously. Many thousands of such beam splitters are used in radar sets to superimpose the image of a cathode ray oscilloscope screen (which might be green) with a map of another color (magenta, for example). It is easy to see an additional color selective beam splitter could be added to the configuration shown in Figure 20.1 to form a three-color separation system which could be used in color photography or color television.[44] Figure 20.91 shows a measured transmission curve of a blue reflector, which is similar to the long-wave pass filter shown in Figure 20.73, but viewed at non-normal incidence. The transmission curves at various angles of incidence also illustrate the angle shift to shorter wavelengths as ϕ increases. Figure 20.92 shows the reflectivity of green reflector at 45° incidence. One method of achieving this narrow reflection band is to use a $3\lambda/4$ stack. This is illustrated in Figure 20.93, which shows the computed transmission of nine layers of zinc sulfide and magnesium fluoride which have an optical thickness of $3\lambda_o/4$ where $\lambda_o = 546$ mμ, matched at 60°. A first order reflectivity peak occurs at 3×546 mμ $= 1.64\ \mu$ in the infrared. The third order high-reflectance zone is centered at 546 mμ. The use of high-order reflectivity peaks is an effective method of achieving a narrow reflection band.

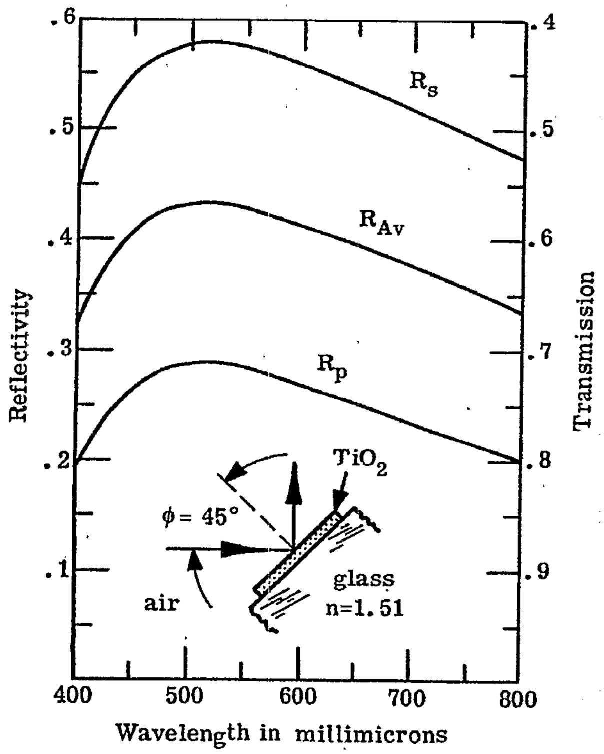


Figure 20.89- Computed R_p, R_s, and R_{av} = 1/2 (R_p + R_s) of a single quarter-wave layer TiO_2 at $\phi = 45$. The dispersion of the index of the TiO_2 is taken into account.

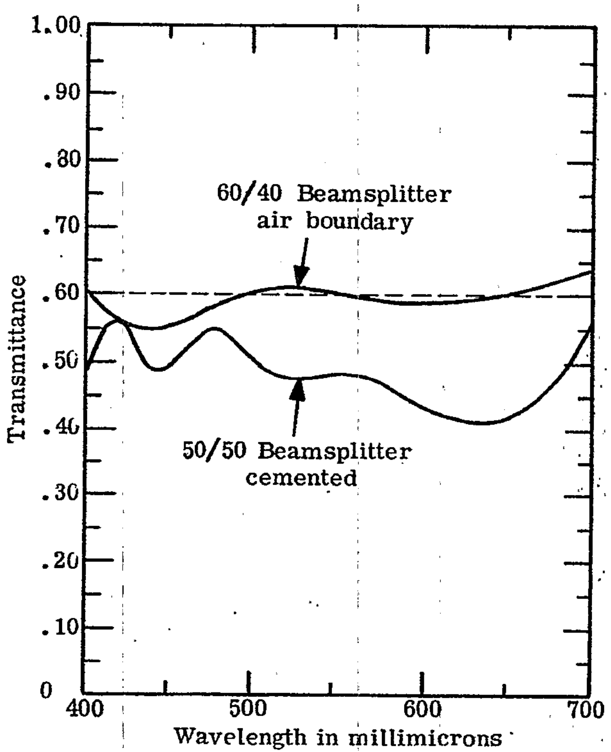


Figure 20.90- Measured transmittance of a 60/40 and a 50/50 multilayer beam splitter at ϕ = 45°. R is nearly 1 = T. Courtesy of Fish-Schurman Corporation.

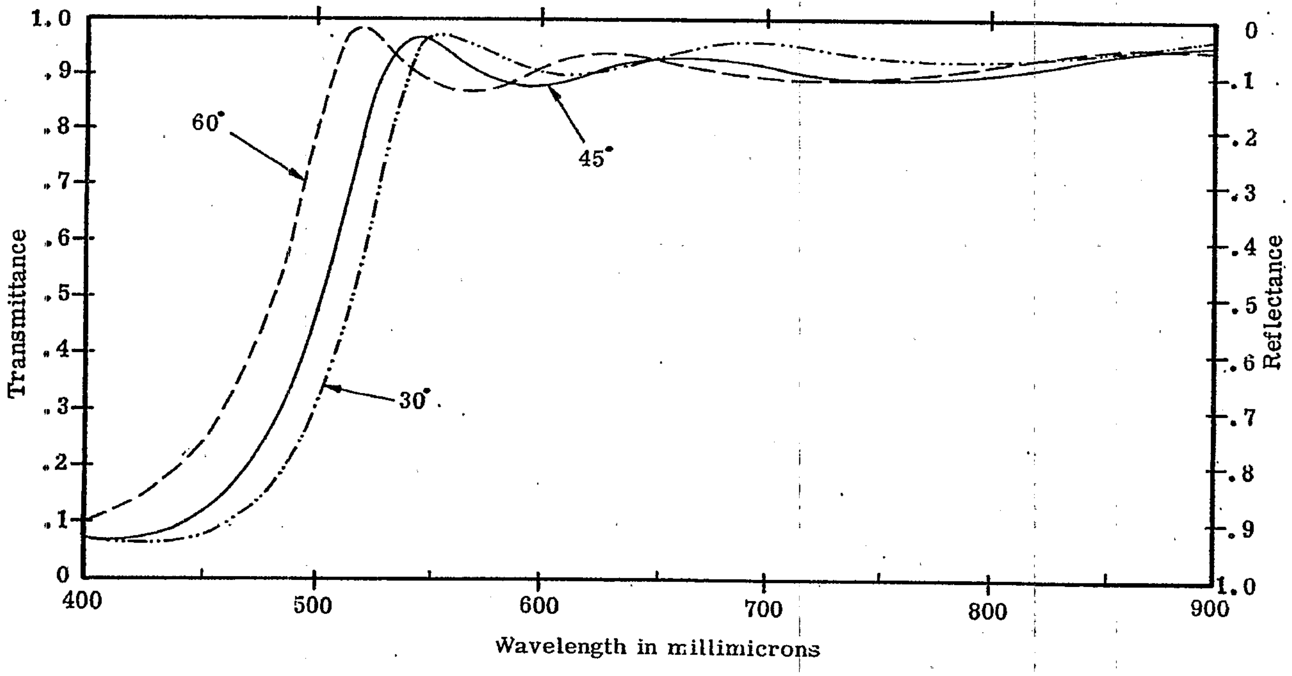


Figure 20.91- Measured spectral transmittance at various ϕ of a color-selective beam splitter. Courtesy of Fish-Schurman Corporation.

20.8 MIRRORS

The distinction between a mirror and a beam splitter is indeed tenuous. Some of the semi-transparent mirrors described in Section 20.8.2 could be used as color selective beam splitters if they were illuminated at non-normal incidence. Also, some of the long-wave and short-wave pass filters described in Sections 20.5 and 20.6 could be equally well used as semi-transparent mirrors. Thus the classification of a multilayer as a long-pass filter, color selective beam splitter, or a semi-transparent mirror depends upon its ultimate use and is not an inherent property of the device itself. Mirrors can be classified as either opaque or semi-transparent. Opaque mirrors are used if one is only concerned with the reflected light, whereas a portion of the light which is not reflected from a semi-transparent mirror, is transmitted.

20.8.1 Opaque mirrors. Silver, aluminum, and rhodium are commonly used as reflecting films in the visible spectral region. All three of these metals can be evaporated in a vacuum. An opaque coating is produced by a film 0.2 μ (eight millionths of an inch) thick. A freshly deposited film of silver has the highest reflectivity, particularly in the red, but the reflectivity deteriorates rapidly in air. An extremely thin, tough film of aluminum oxide forms naturally on the surface of an aluminum coating which protects it against further oxidation. Thus aluminum coatings last a long time under normal environmental conditions. The reflectivity of a rhodium coating is about 10% below that of aluminum. Rhodium is quite inert to attack from salt water and notwithstanding its lower reflectivity it is used in applications where the environmental conditions are severe. Gold films have a high reflectivity in the infrared. Hass[45] has published reflectivity data on the various metal coatings.

20.8.1.2 Protective coatings. A single layer of silicon monoxide(SiO) is often deposited on an aluminum film to protect it from abrasion and chemical attack. The addition of this layer does not alter the reflectivity of the aluminum if the optical thickness of the layer is either much smaller than the wavelength of the incident light or if its optical thickness is a half-wave, in which case it is absentee (see Section 20.1.5.2.2). The former condition can easily be obtained at long wavelengths in the infrared. Figure 21.14 shows the computed spectral reflectivity of aluminum overcoated with a single layer of silicon monoxide. At certain wavelengths the reflectivity of the aluminum is decreased from 0.90 to 0.78 by the addition of the protective overcoat.

20.8.1.3 Reflection enhancing overcoatings. The purpose of adding the single-layer overcoat described in the foregoing paragraph is to increase the resistance of the mirror to abrasion and chemical attack. It is also possible to overcoat a metal mirror with a multilayer coating in order to increase the reflectivity to values as high as .995. Such mirrors are useful in optical systems in which the light is reflected many times. Figure 20.94 shows the computed reflectivity of a bare aluminum mirror. It should be noted that the scale of the ordinate changes at 0.90. The dispersion of the optical constants is taken into account in the calculation; the reflectivity compares well with the published data of Hass[45]. On the same graph is shown the reflectivity of an aluminum mirror overcoated with a six-layer dielectric stack of magnesium fluoride and zinc sulfide. The dispersion of the refractive index of the latter material has also been included in the calculation. The computed reflectivity attains a maximum value of 0.996. Jenkins[46] measured accurately the reflectivity of such overcoated mirrors and found a maximum reflectivity of 0.994. Other types of overcoatings can be used to obtain a broader region of high reflectivity[46], so that the reflectivity does not decrease in the blue, as it does in Figure 20.94.

20.8.2 Semi-transparent mirrors. The mirrors which are discussed in Section 20.8.1 are opaque - that is, the light which is not reflected is absorbed in the metal coating. There are some applications where mirrors are required which not only have a high reflectivity, but also transmit with a high efficiency the light which is not reflected. Such multilayers are useful as coatings for the plates of the Fabry-Perot interferometer and also as coatings for the ends of an optical maser (sometimes called a Laser).

20.8.2.1 Silver films. Silver films have been used for more than six decades as coatings for the Fabry-Perot interferometer. They have the advantage that the single film of silver can be deposited much more quickly and easily than the many films of a multilayer mirror. Kuhn and Wilson[47] measured carefully the R and T of layers of silver on a glass substrate and found that a freshly evaporated film has small absorption loss, particularly in the red spectral region. At a wavelength of $\lambda = 680\ m\mu$, typical values were $R = 0.89$, $T = 0.08$, and the remaining 3% is absorbed. However, in the blue (420 $m\mu$) film with same transmission would absorb more than twice as much. The disadvantage of using silver films is that the reflectivity deteriorates in time. Multilayer coatings have the advantage that they have a lower amount of absorption and that a higher reflectivity can be attained.

20.8.2.2 Dielectric multilayers. Quarter-wave stacks and other types of multilayer coatings have been used quite successfully as semi-transparent mirrors,[46, 48, 49] principally for coating Fabry-Perot interferometers. Figure 20.95 shows the first-order high-reflectance peak of quarter-wave stacks consisting of five, seven, and nine layers of zinc sulfide and cryolite. For a given plate separation, the maximum resolution of a Fabry-Perot is limited by the flatness of the interferometer plates and increasing the reflec-

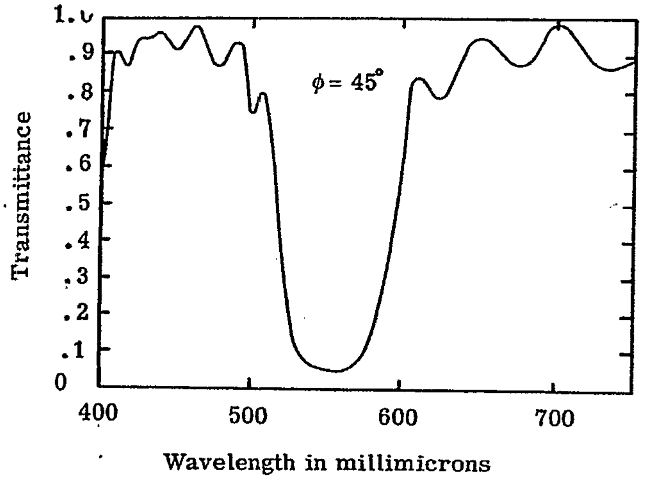


Figure 20.92- Measured spectral transmittance at $\phi = 45°$ of a color selective beam splitter which reflects the green. Courtesy of Fish-Schurman, Inc.

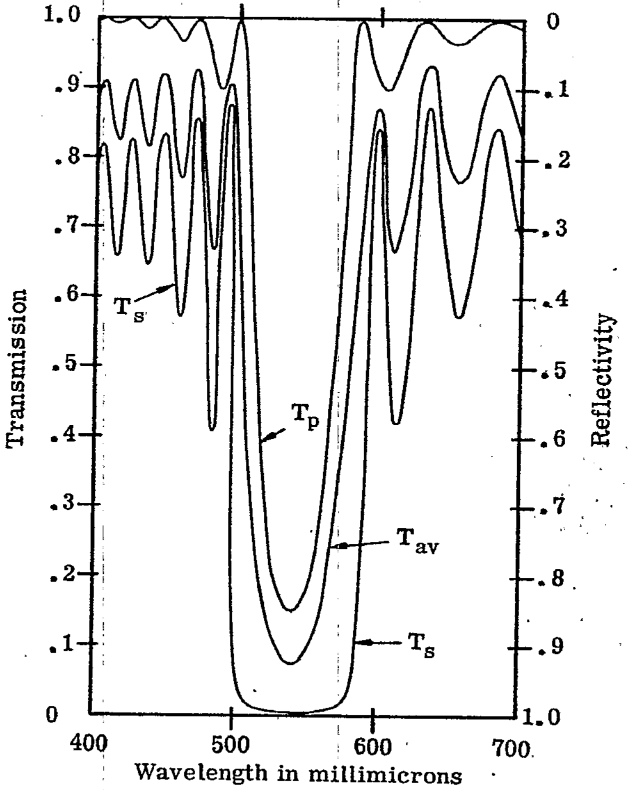


Figure 20.93- Computed T_p, T_s, and T_{av} = $1/2\ (T_p + T_s)$ of a color-selective beam splitter at $\phi = 65°$. The design is glass $H\ (LH)^4$ air, $n_L = 1.35$, $n_H = 2.30$, $n_L t_L = n_H t_H = \frac{3\lambda_o}{4}$ at $\phi = 65°$, for $\lambda_o = 550\ m\mu$.

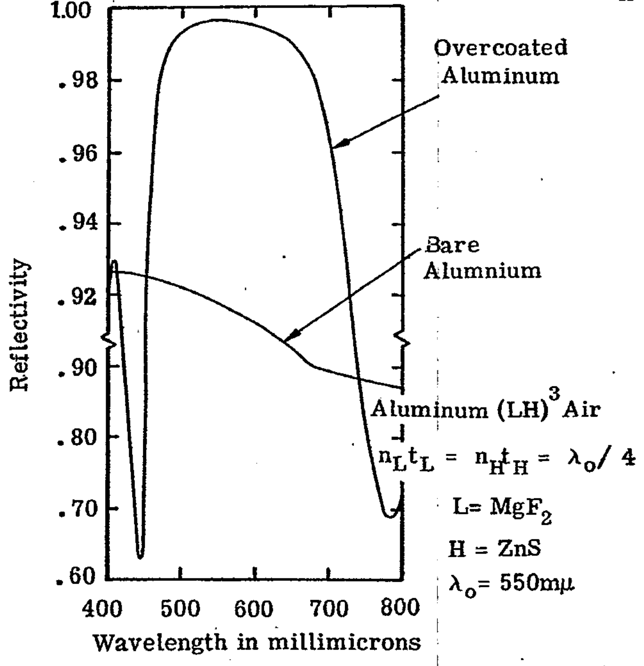


Figure 20.94- Computed spectral reflectivity of bare aluminum and aluminum overcoated with six dielectric layers. The scale of the ordinate changes at 0.90. The disperson of the optical constants is accounted for.

tivity over a certain value merely results in a loss of light without any gain in resolving power. Hence the seven-layer films are adequate for most plates. The advantage of using dielectric coatings over silver coatings is that the absorption loss of the former is quite low. A typical multilayer dielectric stack has an absorption of 0.005.[50] The disadvantage of using quarter-wave stack is readily apparent from Figures 20.71 or 20.95; the region of high reflectivity covers only about one-half of the visible spectrum. This difficulty is avoided if a broad-band reflector[49] is used, such as the one whose spectral reflectivity is shown as curve "B" in Figure 20.95. The design of this reflector is given in reference 49. The reflectance of this multilayer is close to 0.95 in the visible spectrum, which is the optimum value for most interferometer plates. The reflectivity of the cold mirror shown in Figure 20.74 is too high and would not be suitable for coating most interferometer plates.

20.8.2.3 From a theoretical point of view, (i.e. Equations (40) - (43)) the reflectivity of a quarter-wave stack can be made as close to 1.0 as we choose by simply using a sufficient number of layers in the stack. At the present state of thin film technology, it is found that the reflectivity of a $ZnS - MgF_2$ stack does not increase appreciably beyond fifteen layers. The fact which limits the ultimate reflectivity is that the layers are composed of small crystallites which are randomly oriented. This means that the layers have random fluctuations in their optical thickness and also that there are irregularities at the surfaces of the layers which scatter light. Giacomo[51, 52] has studied this problem in detail. As is discussed in 20.10.6.1, it is this scattering which limits the band width of dielectric interference filters. It is possible that in the future techniques will be developed to reduce this scattering, thereby making it possible to increase the reflectivity to values very close to 1.0 and consequently to produce multilayer interference filters with extremely narrow band widths.

20.9 BAND PASS FILTERS

A band pass filter is a multilayer which has a high transmission in a specific spectral region and attenuates in both the short and long wave regions on either side of the pass band. In a certain sense, all of the short-wave pass filters described in 20.6 are band pass filters, because the higher order high-reflectivity peaks limit the width of the short-pass region, although this is not the intent of such a filter. Included in the class of band pass filters are Fabry-Perot type filters or interference filters. This filter, which is discussed in 20.10, was the first type of filter whose transmission characteristics depended upon the interference of light, rather than the absorption of light. The name interference filter is retained for primarily historical reasons. Strictly speaking, every multilayer filter, from the single-layer coating to a sixty-layer stack, is an interference filter, in the sense that its transmission characteristics depends upon the interference of the light reflected from various layers within the filter. For reasons described in 20.10, we prefer to call this type of filter a Fabry-Perot type filter.

20.9.1 Filters with a wide pass band. From the discussion of long-wave and short-wave pass filters in 20.5 and 20.6, it is evident that a band pass filter can be constructed simply by combining a long-pass and short-pass filter. They can be either deposited on the same substrate or they can be deposited on separate substrates as a composite filter. The advantage of constructing a band pass filter in this manner is that width and position of the pass band can be changed at will by altering the cutoff of either the short or long-wave stack. Also, the amount of attenuation outside of the pass region can be easily controlled by changing the number of layers in either of the stacks. The width of the pass band is essentially independent of the attenuation outside the pass region. This is not true of the Fabry-Perot type filter described in 20.10. There are also other possibilities, such as combining a short-wave pass multilayer filter with a glass or dye absorption type filter which passes in the long-wave region. Figure 20.95 shows the spectral transmission of some assorted filters with a broad pass band in the visible and near ultraviolet spectral regions, while Figure 20.96 shows some similar filters for the infrared.

20.10 FABRY-PEROT TYPE FILTERS (INTERFERENCE FILTERS)

20.10.1 Basic concepts - the Fabry-Perot interferometer. In order to explain how an interference filter functions, it is first useful to understand how a Fabry-Perot interferometer works and to become familiar with such concepts as the "Q", free spectral range, order number, and so on. The basic idea of the interference filter has been known since the turn of the century, when two French physicists, C. Fabry and A. Perot, invented the interferometer which bears their name. Although this device is used mostly to measure fine details of emission line spectra, it can also be used as a filter with a very narrow pass band. The Fabry-Perot interferometer, which is shown in Figure 20.98, consists of two plates of glass or fused quartz which have been polished to a high degree of flatness. The plates are held apart by three separator pins so that the inner faces of the plates are parallel. These faces are coated with a semi-transparent mirror, such as the type described in 20.8.2. The transmission T_f of the interferometer is derived in most books on physical optics[53, 54] and in Section 5.17.

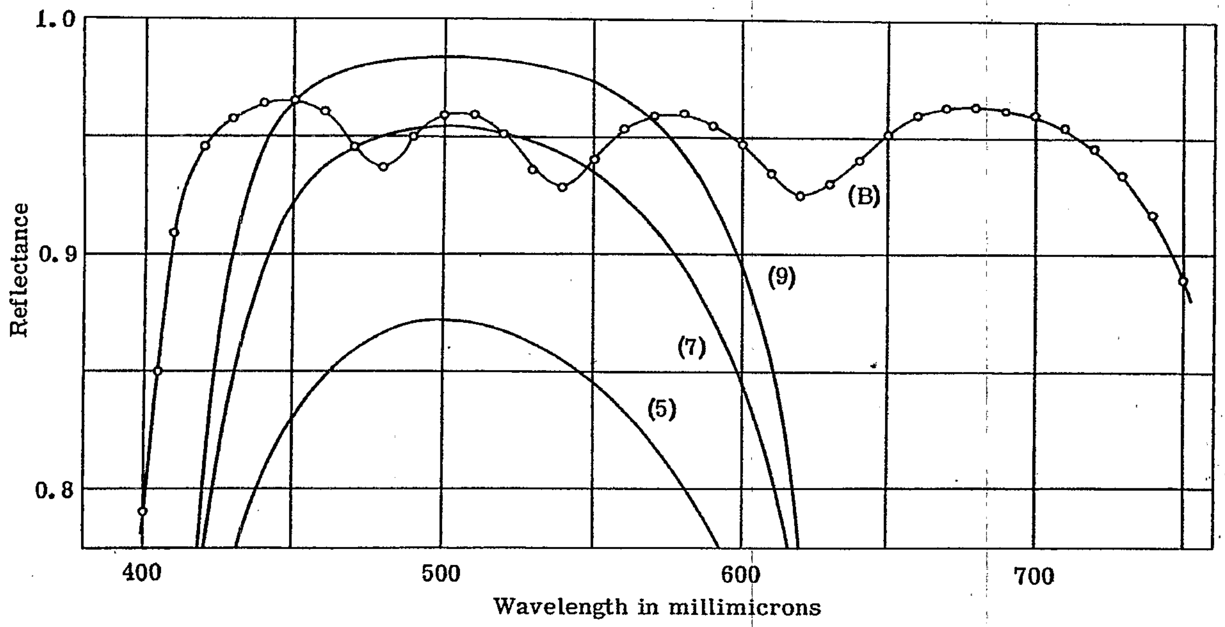


Figure 20.95- Measured spectral reflectivity of quarter-wave stacks with 5, 7, and 9 layers and also a broad-band reflector consisting of 15 layers of ZnS and cryolite. From ref. 49.

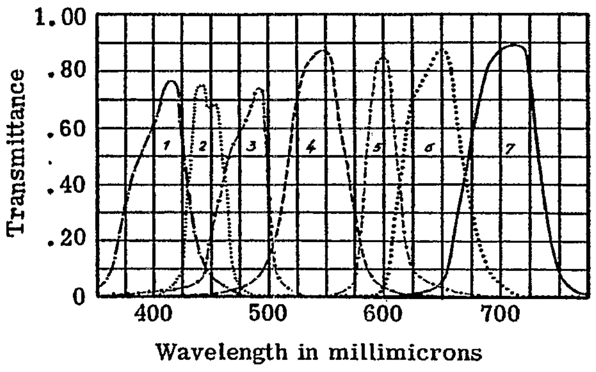


Figure 20.96- Measured transmittance of band pass filters which utilize both multilayer and glass absorption filters. Courtesy of Balzers Aktiengesellschaft, Liechtenstein.

MEASURED TRANSMITTANCE OF A SERIES OF BAND PASS FILTERS COMPLETE WITH AUXILIARIES

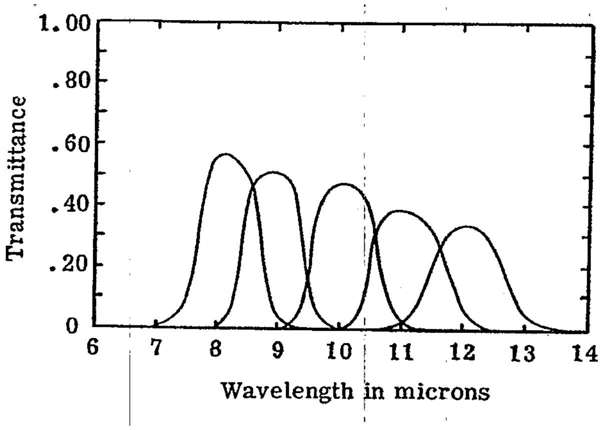


Figure 20.97- Measured spectral transmittance of some infrared band pass filters, complete with auxilliaries. Courtesy of Bausch and Lomb, Inc.

$$T_f = T_{max} \left(1 + F \sin^2 \eta\right)^{-1} \tag{49}$$

where

$$T_{max} = \frac{T_1 \; T_2}{(1 - R)^2} , \tag{50}$$

$$F = \frac{4R}{(1 - R)^2} , \tag{51}$$

$$R = \sqrt{R_1 R_2} , \tag{52}$$

and

$$\eta = 2\pi\sigma n_g t_g - \frac{\epsilon_1 + \epsilon_2}{2} . \tag{53}$$

T_1 , T_2 and R_1 R_2 are the reflectivity (measured on the side of the mirror coating facing the gap) and transmission of the semitransparent mirror coatings on each of the plates. The coatings need not have the same R and T, and can also be absorbing, so that $R + T \neq 1$. ϵ_1 and ϵ_2 are the phase shift upon reflection from each of the coatings, which is explained in Section 20.10.1.4. t_g is the physical separation of the plates, measured from the surface of the coatings and n_g is the refractive index of the medium in the gap, which is usually air or a vacuum. In the most general case, R_1, T_1, ϵ_1, R_2, T_2, and ϵ_2 all vary with wavelength.

Since the $\sin^2$ function in the denominator of Equation (49) is always a positive quantity, T_f attains a maximum value when $\eta = m\pi$; the integer m is called the order number. If we neglect for the moment the phase shifts, ϵ_1 and ϵ_2, in Equation (53), then T_f is a maximum when

$$m\pi = 2\pi\sigma n_g t_g , \tag{54}$$

or

$$m \frac{\lambda_o}{2} = n_g t_g . \tag{55}$$

Equation (55) states that the resonant, or maximum transmission condition, occurs when the separation of the plates is an integral number of half-waves. The Fabry-Perot interferometer can be considered as a resonant cavity for light waves. Similar resonance conditions occur when an integral number of half-waves fit into a microwave cavity or an organ pipe. One difference is that the resonance in microwave cavities is usually in the first or second mode, that is, $m = 1$ or 2. A very high mode - or to use the language of physical optics, a high order of interference - is obtainable in the Fabry-Perot interferometer. For example, a plate separation of $t_g = 1.0$ cm. is commonly used, in which case $m = 40{,}000$ at $\lambda = 500$ mμ ; that is, 40,000 half-waves fit in between the plates.

20.10.1.2 Spectral transmission of F.P. interferometer. If the Fabry-Perot interferometer is illuminated with highly collimated light, as shown in Figure 20.99, then its transmission, T_f , as a function of wave number (frequency) consists of a series of transmission peaks which are equally spaced, as is shown in Figure 20.100. The spacing σ_f between adjacent transmission peaks is called the free spectral range of the interferometer and is inversely proportional to the separation of the plates:

$$\sigma_f = 1/2t_g \tag{56}$$

For most plate separations which are attainable, this free spectral range is rather small and hence many transmission peaks are packed close together. For example, suppose that two plates are separated by three thin pieces of metal foil which are a little less than one thousandth of an inch thick, so that $t_g = 0.0025$ cm. From Equation (56), we find that $\sigma_f = 200$ cm^{-1} . This means that in the green region of the spectrum, transmission maxima would occur at wave numbers of 20,000 cm^{-1} , 19,800 cm^{-1} , 19,600 cm^{-1} , and so on. Converting these wave numbers into wavelength (see 20.1.3.3), we find that transmission peaks occur at 500.00 mμ , 505.05 mμ , 510.10 mμ , and so on. Thus the separation of the peaks is about 5 mμ in this spectral region. If this interferometer were used to pass a spectrum line at 505.05 mμ , then it would be necessary to have an auxiliary blocking filter (a composite filter, to use the language of 20.5.1.3.2) to attenuate the unwanted transmission peaks at 500mμ , 510mμ , and at other wavelengths.

Figure 20.101 shows the shape of the transmission band of an F.P. interferometer whose plates are coated with highly-reflecting films so that $R_1 = R_2 = 0.95$. This is merely an enlarged view of one of the transmission bands shown in Figure 20.100, translated to a wavelength scale. The maximum transmission of the band is T_{max} and occurs at a wavelength λ_o. By definition, the total width of the band at $1/2\ T_{max}$ is $\Delta\lambda_{1/2}$ *. Thus $\Delta\lambda_{1/2}$ gives an indication of the narrowness, or the sharpness, of the resonance line of the interferometer, and consequently its ability to transmit radiant energy at its resonant wavelength, λ_o, and to attenuate radiant energy of nearby wavelengths. This is allied to the chromatic resolving power of the interferometer.

20.10.1.3 Resolving power of the F.P. interferometer. The narrowness of the resonance line in the frequency spectrum of any resonant device is proportional to the ability of the device to store energy with low loss of energy per cycle of oscillation. The dimensionless quantity "Q" of a resonant system is defined as:[55a]

$$Q = 2\pi \frac{\text{Energy stored in system}}{\text{Energy lost in one cycle}}, \qquad (57)$$

and is inversely proportional to the line width, $\Delta\sigma_{1/2}$. For example, the simple "LC" resonant circuit shown in Figure 20.101 stores energy in the electric field in the capacitor and in the magnetic field of the inductor. Each cycle a fraction of this energy is dissipated in the form of heat due to the finite resistance of the wire in the inductor and to losses in the dielectric of the capacitor. If this loss can be decreased, then the Q of the circuit will improve and its sharpness as a tuning element, say in a radio receiver, will improve. Similarly, the energy of a microwave cavity is usually extracted through a small hole in the wall of the cavity. During each cycle of oscillation, a small amount of the energy which is stored in the electromagnetic fields in the cavity leaks out of this hole and is also absorbed as Joule heat in the walls of the cavity. The Q of the cavity is improved by silver plating the cavity walls and thus reducing this loss. The case for the F.P. interferometer is completely analogous. Standing light waves are established in the gap between the two reflecting plates, and electromagnetic energy is stored in this gap. During each cycle of oscillation, which in the case of visible light is more than 10^{14} cycles per second, a small amount of this energy in the cavity is depleted by either being absorbed or transmitted through the coatings.

From the foregoing considerations, we would expect that the Q of a F.P. interferometer would increase as the reflectivity of the coatings on the plates is enhanced. Another method of improving the Q is to store more energy in the cavity, which is accomplished by simply making cavity larger - that is, by using a larger separation of the plates. A more quantitative analysis,[55b] such as the one in Section 5.16, leads to the relation

$$Q = \left(\frac{\lambda_o}{\Delta\lambda_{1/2}}\right)\left(\frac{\sigma_o}{\Delta\sigma_{1/2}}\right) = \left(\frac{\sqrt{R}}{1 - R}\right) m\pi . \qquad (58)$$

Thus an interferometer with $R = 0.95$ and a plate separation of 1 cm. has a Q over two million, which is larger by several orders of magnitude than any microwave device. The Q is a quite sensitive function of R, as R gets close to unity. The Q computed from Equation (58) can either be increased or decreased by the variation with wavelength of the phase shift on reflection of the two semi-transparent mirrors. A more exact formula is given in reference 56. It should be remarked that Q is used as a criterion of the narrowness of the pass band in preference to the chromatic resolving power, which is used in most books on physical optics. The reason is that there are many different criteria for defining the chromatic resolving power and it is not desirable that Equation (58) be confused with them.**

20.10.1.4 The phase shift upon reflection. A standing wave is created when a light wave is incident upon a reflecting surface. The positions of the nodes of this standing wave remain fixed in space. Suppose an incident light wave, whose electric vector is represented by

$$E = E_o\, e^{i(\omega t - 2\pi z/\lambda)} \qquad (59)$$

---

* If the width $\Delta\lambda_{1/2}$ is translated into the corresponding width, $\Delta\sigma_{1/2}$, on a frequency (wave number) scale, they are related by: $\frac{\Delta\lambda_{1/2}}{\lambda_o} = \frac{\Delta\sigma_{1/2}}{\sigma_o}$, were $\lambda_o = \sigma_o^{-1}$

** For example, compare Eq. 58 in this text with Eq. 43 on page 334 of Born and Wolf[11].

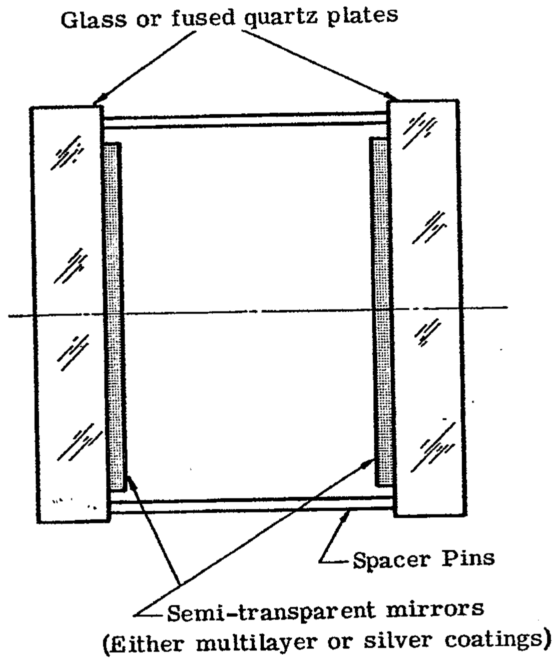

Glass or fused quartz plates
Spacer Pins
Semi-transparent mirrors
(Either multilayer or silver coatings)

Figure 20.98- The essential parts of a Fabry-Perot interferometer.

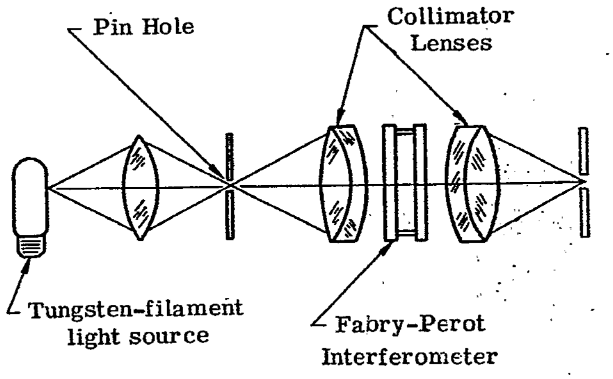


Figure 20.99- A method of illuminating the Fabry-Perot interferometer with collimated light so that it can be used as an optical filter.

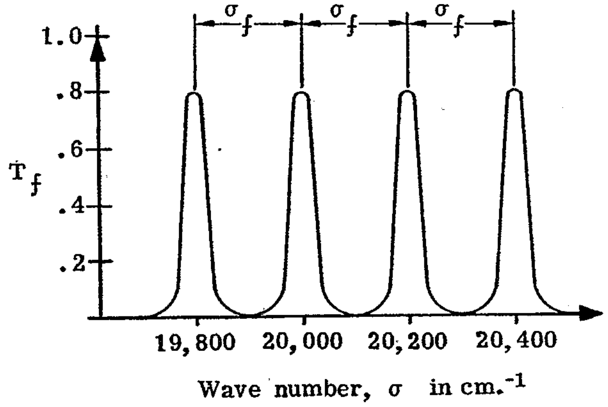
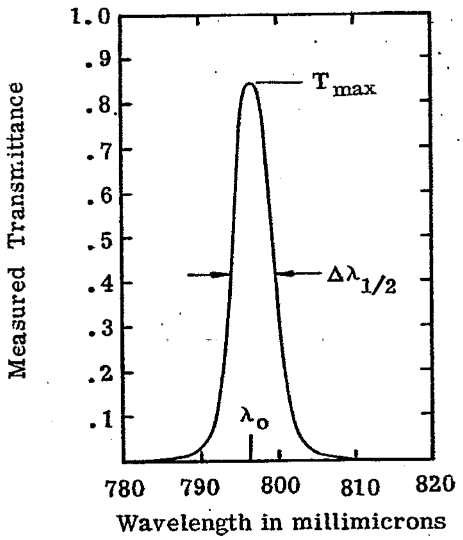


Figure 20.100 – The transmission as a function of wave number of a Fabry-Perot interferometer with t_g = 0.0025 cm .



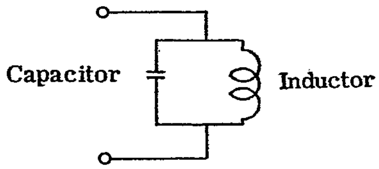


Figure 20.101- Measured spectral transmittance of Fabry-Perot type multilayer filter. Courtesy of Dr. C. Alley.
A parallel resonant circuit.

is reflected from a surface. If the reflected wave represented by:

$$E_r = R^{1/2} E_o e^{i(\omega t + \epsilon_1 + 2\pi z/\lambda)} , \qquad (60)$$

this defines the phase shift upon reflection, ϵ_1 . If the reflecting surface is a metal with infinite conductivity - that is, a perfect reflector, - then the standing wave has a node at the surface and the phase shift, ϵ_1, is 180°. At some instant of time, the amplitude of the light wave varies sinusoidally in space, as shown by the solid line in Figure 20.102; a half-cycle later it is represented by the dashed line. If ϵ is less than 180°, as would be the actual case for a silver mirror, then the node of the standing wave is to the right of the reflecting surface. If ϵ is greater than 180°, it is to the left, as shown in Figure 20.102. The phase shift upon reflection is exactly 180° for an air-glass interface, and does not vary with wavelength as long as the glass is optically non-absorbing. The ϵ for silver films changes slowly with wavelength, but this change is so small that it is usually neglected. The phase shift upon reflection for multilayer is much larger and, as is shown in 20.10.5, modifies the shape of the transmission band of a multilayer filter. It can also alter the wavelength at which the pass band occurs.[56]

20.10.1.5 Effect of ϵ on a resonant cavity. The effect of the phase shift upon reflection is to shrink or expand the walls of a resonant cavity. If the variable η in Equation (53) were defined as $\eta = 2\pi\sigma n_g t_g$, thus omitting the second term, this would tacitly assume that there is a node of the standing wave at each of the two reflecting surfaces. The addition term in Equation (53), $1/2\,(\epsilon_1 + \epsilon_2)$ can be regarded as a correction term which accounts for the fact that the ϵ of each of the mirrors shifts the node of the standing wave. This shift in the node of the standing wave changes the resonant frequency of the cavity. If ϵ_1 and ϵ_2 change with wavelength, this alters the shape of the transmission band from the Lorentzian shape shown in Figure 20.101.

20.10.1.6 Shape of the transmission band. If the condition is fulfilled that ϵ_1 and ϵ_2 are constant with wavelength and that R_1 and R_2 are large enough so that F is much greater than one, then over the narrow wave number range near a pass band at σ_o , the sine function in Equation (49) can be replaced by its argument and T_f written as

$$T_f = (T_{max})\,(1 + G\,\Delta^2)^{-1} \quad \text{where} \quad \Delta + \sigma_o = \sigma \qquad (59)$$

and constants such as F , 2π , t_g , etc. have been lumped together into G , which is essentially constant over the range of Δ, in which the approximation is valid . The point we wish to make is that this is a Lorentzian line shape (shown in Figure 20.101), and is similar to the shape power absorption curve of a series resonant "LC" circuit with a high Q[57] . A characteristic of the Lorentzian line shape is that a long tail which decreases slowly in amplitude extends to both the short and long-wave side. There are applications of F.P. type filters which require that the transmission on either the short-wave or long-wave side, decrease much more rapidly than is provided by a filter with a Lorentzian-shape transmission band. A few multilayers with a non-Lorentzian transmission band are shown in 20.10.7.

20.10.2 Fabry-Perot type multilayer filters. The Fabry-Perot interferometer described in the foregoing section is actually used as a band-pass filter in the laboratory. Notwithstanding the high Q which can be attained, it is not widely used because the optically polished plates are quite expensive and small mechanical vibrations or temperature changes cause the plates to warp out of parallel, thus degrading the Q of the instrument. These difficulties are partially avoided if the material in between the plates (i.e. the spacer) is a solid material. This is accomplished by depositing a mirror on a substrate by evaporation in a vacuum, then evaporating a spacer layer whose optical thickness satisfies Equation (55), and then evaporating another mirror. Another method is to use a thin piece of mica as the spacer material and to evaporate a mirror on both sides. This technique is discussed in 20.10.5.2.

20.10.2.1 Method of analysis. In the case of the Fabry-Perot interferometer, it is quite natural to analyze the performance in terms of Equation (49). This has the advantage that one does not need to know any of the details about the construction of the semitransparent mirrors on the two faces - it is only necessary to know the R, T, and ϵ of each mirror, and n_g , t_g . However, in the case where the spacer is evaporated as an integral part of the filter, such as the filters depicted in Figures 20.103 or 20.108, then there are many methods of analyzing its spectral transmission, as for example the admittance method[58] , the matrix method[59] or by considering it as a Fabry-Perot interferometer[60] . There is a large class of multilayer filters whose spectral transmission is most easily analyzed by treating them like a Fabry-Perot interferometer. This is accomplished by selecting one of the films in the stack - usually the center film if the stack has an odd number of layers and is symmetrical - and considering this film as a spacer layer with an optical thickness $n_g t_g$. The remaining films in the stack are divided into two groups - those between the spacer layer and the substrate form an effective interface "A" and the films between the spacer layer and the incident medium form the other effective interface.[60] The spacer layer is usually, but not

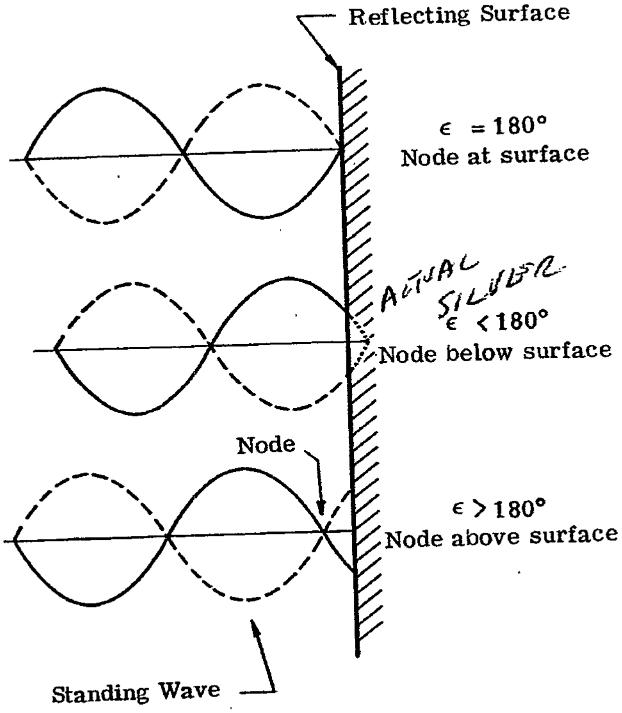


Figure 20.102- Showing the node of the standing wave as a function of the phase shift upon reflection, ϵ .

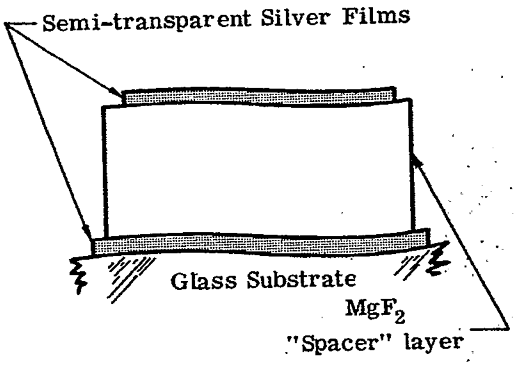
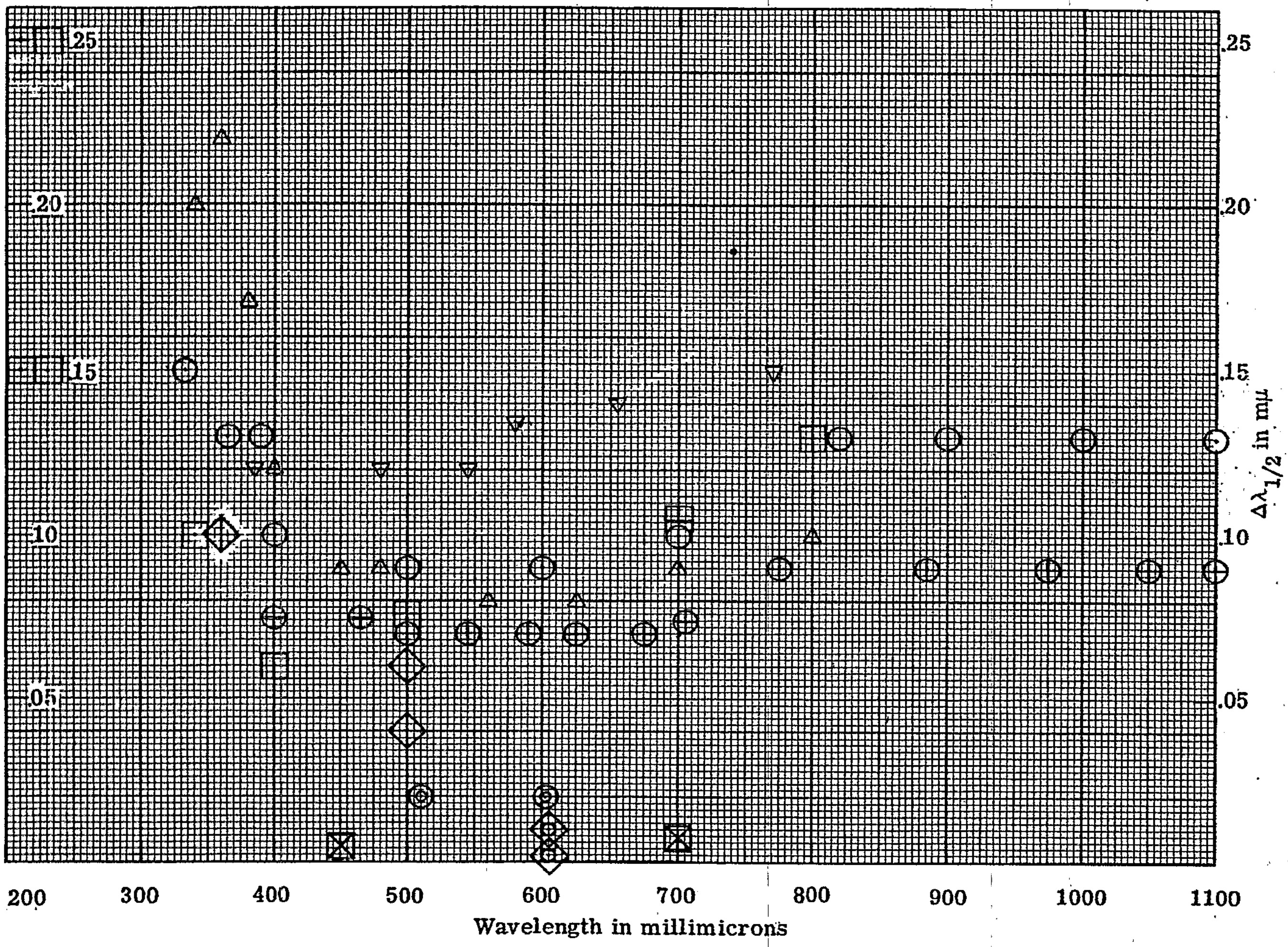


Figure 20.103- An enlarged view of a silver-dielectric-silver Fabry-Perot type filter, the lack of flatness of the substrate being exaggerated for purposes of illustration.



LEGEND: Filters contain metal unless specified otherwise.

☒ Baird-Atomic, Inc. Multilayer dielectric with Q = 1000. $T_{max.}$ = 0.45 to 0.60, blocking filters included.

⊞ Baird-Atomic, Inc. Multilayer dieletric filter with rectangularly shaped pass band. $T_{max.} > 0.60$, blocking filters included.

⊡ Baird-Atomic, Inc. Type "A"

$T_{max.}$			$\Delta\lambda_{1/2}$
0.15	→	0.20	25 mμ
0.10	→	0.15	15 mμ
0.05	→	0.10	10 mμ

△ Bausch and Lomb, Inc. "Standard series", second order interference, $T_{max.}$ = 0.30 to 0.35.

⊙ Carl Zeiss, Jena$^{\underline{er}}$, "Single filters", $T_{max.}$ = 0.25 in the u.v. and near i.r. . $T_{max.}$ = 0.35 in the visible.

⊕ Carl Zeiss, Jena$^{\underline{er}}$, "Double filter" (a composite filter) Type DSIF, $T_{max.}$ = 0.08 to 0.15.

⟐ Multilayer dielectric, $T_{max.}$ = 0.93, without blocking filters. See reference 64.

◈ Multilayer dielectric, consisting of from 21 layers ($T_{max.}$ = 0.55) to 29 layers ($T_{max.}$ = 0.15). See reference 65.

∇ Schott (Jena$^{\underline{er}}$ Glaswerke Schott & Gen., Mainz) These are "Line filters" of unspecified order of interference.

◎ Specially prepared F-P type filter consisting of nine dielectric films and two silver films. $T_{max.}$ ~ 0.40. See reference 64.

◇ Spectrolab, Inc. Published curves show $T_{max.}$ ~ 0.90.

Figure 20.104- The total width at .5 $T_{max.}$, $\Delta\lambda_{1/2}$, and wavelength of the passband, λ_0, of some representative Fabry-Perot type filters composed of the metal-dielectric-metal type and all-dielectric type. See 20.10.2.2 for details.

always, an integral number of half-waves in optical thickness. References 60 and 61 show some exceptions. Thus by dividing a multilayer into a spacer layer and two effective interfaces, it can be analyzed as a Fabry-Perot interferometer and Equation (49) can be used to compute its transmission, provided R_1, T_1, ϵ_1, R_2, T_2, ϵ_2 for the interfaces are known.

The simplest type of multilayer whose performance can be analyzed as a F.P. interferometer is the silver-dielectric-silver interference filter (shown in Figure 20.103). When its discovery was announced, it was called an interference filter, since it was the first type of filter which operated on the basis of the interference of light, rather than absorption. When other types of multilayers came into use a decade later, it was recognized that all multilayers are in a sense "interference" filters, since their transmission characteristics depend upon the interference of light reflected from various layers within the multilayer. Thus in this section, the term "Fabry-Perot type filter" (F-P type) is used in preference to "interference filter".

20.10.2.2 Criteria for evaluating F-P type filters. Most F-P type filters are band pass filters and have a narrow transmission spike in the pass region and a high attenuation outside of that region. In order to compare the performance of different types of F-P filters, the following attributes are sometimes considered:

(1) The wavelength λ_o of the maximum transmission of the pass band.

(2) The maximum transmission of the pass band, T_{max}.

(3) The total width of the band at half intensity (i.e. at 1/2 T_{max}) $\Delta\lambda_{1/2}$. This is related to Q: $\Delta\lambda_{1/2} = \lambda_o/Q$.

(4) The Q of the filter, or alternatively, Q^{-1}.

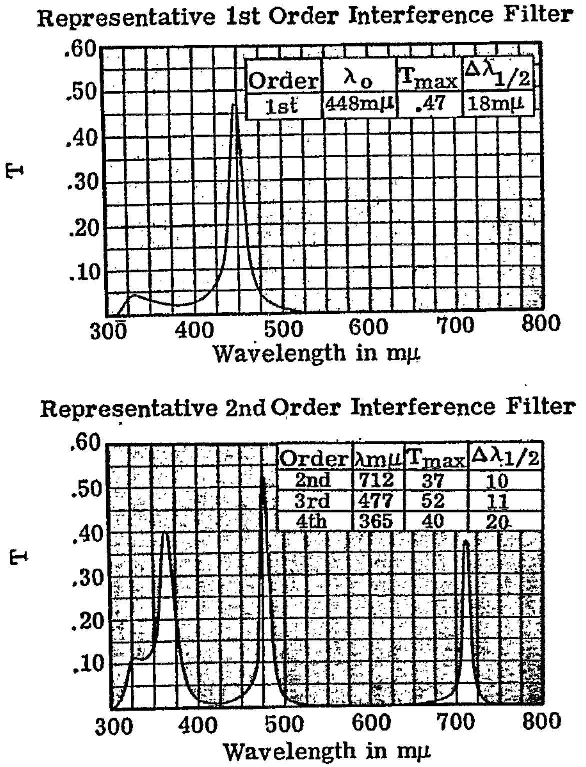
(5) The extent of the attenuation region, and whether blocking filters need be added to extend this region.

(6) The shift of the transmission band as a function of the angle of incidence.

In comparing the literature of various manufacturers of multilayer F-P type filters, it is evident that even if all of these data are given, there is still no substitute for a spectral transmission curve. $\Delta\lambda_{1/2}$ is not necessarily a good criterion for comparing filters, because different types of filters have transmission bands of different shapes. For example, the pass band of the silver-dielectric-silver F-P type filter shown in Figure 20.105 has a Lorentzian line shape and shows the characteristic long tail of high transmission towards the blue, whereas the filter in Figure 20.118 has a pass band of a different shape and although its $\Delta\lambda_{1/2}$ is twice as large, the transmission in the blue decreases much faster. Sometimes the T_{max} given in the specifications of a manufacturer may or may not include the blocking filters which should be placed in series with the F-P type filter in order to eliminate unwanted transmission bands, as appear in the filters shown in Figures 20.105 and 20.115. Figure 20.114 shows how such a blocking filter is used to eliminate an unwanted transmission band at short wavelengths. The $\Delta\lambda_{1/2}$ at various λ_o of F-P type band-pass filters is shown in Figure 20.104. These data are compiled from the scientific literature and from the catalogues of manufacturers. In the latter case, they represent some of the filters produced by some manufacturers. This is definitely not a comprehensive list, but rather it is intended to give some idea of the range of the $\Delta\lambda_{1/2}$ and T_{max} which can be achieved at various wavelengths. Neither is this list intended to be encyclopedic, including all manufacturers. As was mentioned previously, the parameters T_{max} and $\Delta\lambda_{1/2}$ often do not adequately describe the performance of a filter, and hence it would be incorrect to conclude from Figure 20.104 that manufacturer "A's" filters are superior to "B's" filters. Evidently there is some variation in both the T_{max} and $\Delta\lambda_{1/2}$ between individual filters of a given type, because most manufacturers gave a range of values. Average values are used for the data shown in Figure 20.104

20.10.3 Band pass filters containing metal films.

20.10.3.1 The metal-dielectric-metal (M-D-M) Fabry-Perot type filter. The simplest filter of the F-P type is a three-layer filter consisting of a dielectric film, such as magnesium fluoride, sandwiched between two semitransparent metal films. In 1939 Geffcken[62, 63] applied for a patent on the device in which both the metal films and the dielectric spacer layer film are evaporated. The beauty and simplicity of this method is that such a filter can be deposited on a substrate of common window glass, rather than an extremely precise optical flat, as is requisite for the F-P interferometer. The point is, that the surface of a piece of window glass deviates many wavelengths from being optically flat, but that the three or more layers in the multilayer coatings follow the contour of the substrate. The drawing in Figure 20.103 shows this effect; the lack of planeness in the substrate is greatly exaggerated. Actually, the lack of planeness in the substrate does broaden the transmission bands slightly, but this effect is certainly not noticeable with broad 10 mμ band widths typical of this type of filter.



Order	λ_o	T_{max}	$\Delta\lambda_{1/2}$
1st	448mμ	.47	18mμ

Order	λmμ	T_{max}	$\Delta\lambda_{1/2}$
2nd	712	37	10
3rd	477	52	11
4th	365	40	20

Figure 20.105- Measured transmittance of some metal-dielectric-metal Fabry-Perot type filters. Courtesy of Bausch and Lomb, Inc.

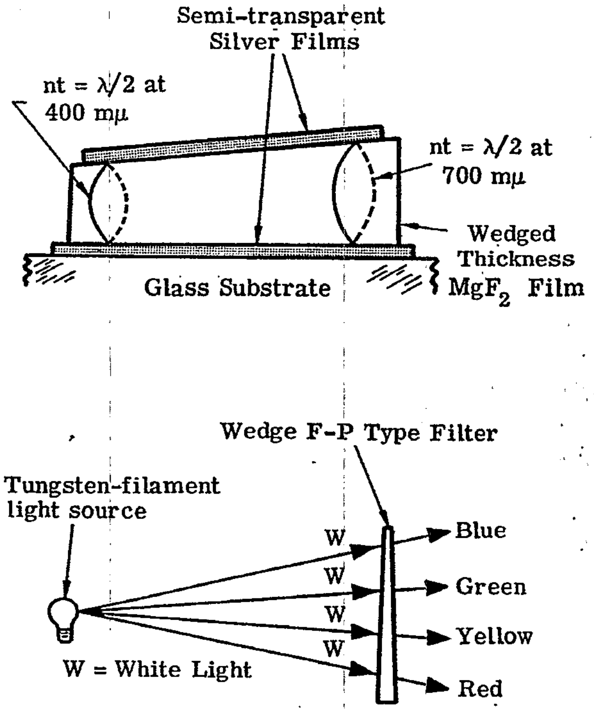


Figure 20.107- (Upper) A cross section of a wedge Fabry-Perot type filter. (Lower) Showing the use of this type of filter as a monochromator.

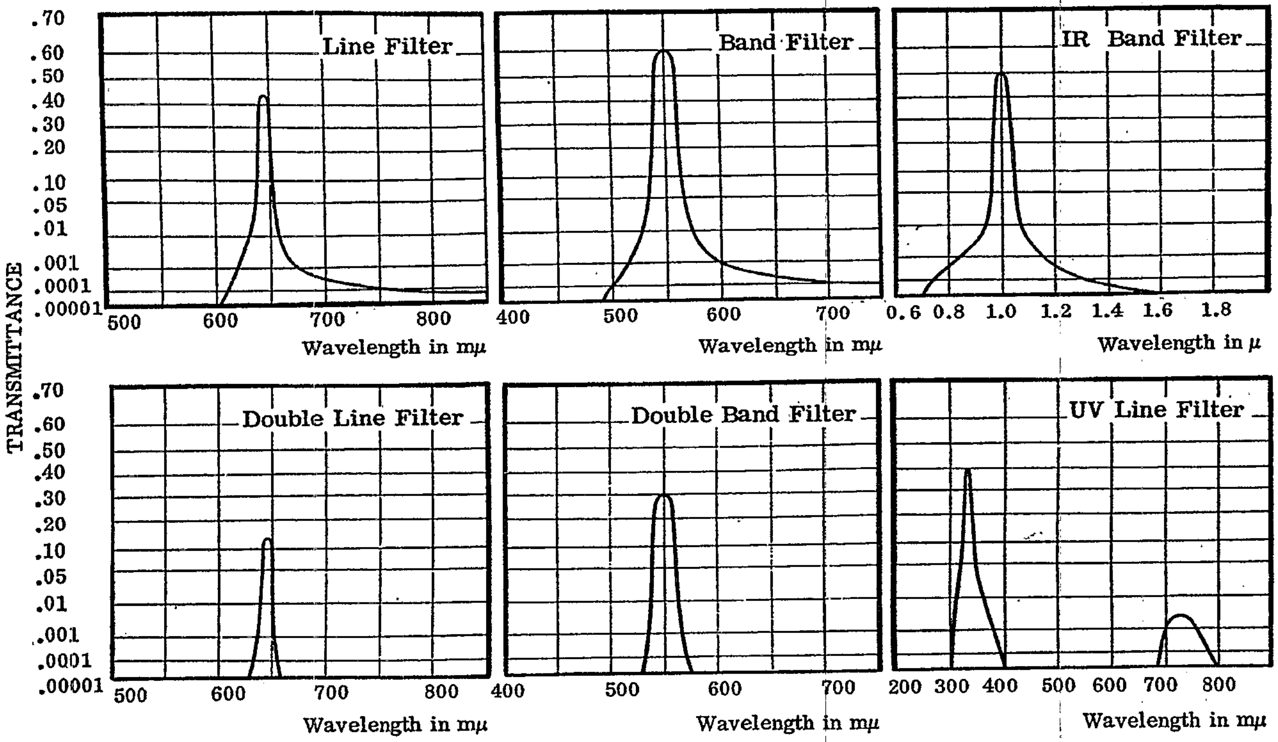


Figure 20.106- Measured transmittance (on a logarithmic scale) of silver-dielectric-silver Fabry-Perot type filters. Courtesy of Schott & Gen., Mainz, West Germany.

For wavelengths greater than 340 mμ, silver is the best metal film to use, having the highest reflectivity and lowest amount of absorption. Below 340 mμ the optical constants of silver change rapidly and the reflectivity drops to very low values. This is why silver-dielectric-silver F-P type filters have a transmission "leak" in this region, as is shown in Figure 20.105. At these shorter wavelengths aluminum is generally used as the metal film, although it absorbs a large fraction of the radiant energy and consequently these filters usually have a T_{max} of .10 to 20. Figure 20.105 shows the spectral transmittance of a silver-dielectric-silver F-P type filter with a first order transmission maximum at 450 mμ. The transmission "leak" in the ultraviolet is due to the loss in the reflectivity of the silver in this region. Also shown in Figure 20.105 is the transmittance of a filter which has a second order peak at 712 mμ, and third and fourth order peaks at shorter wavelengths. In Figure 20.106 is depicted the transmittance (on a logarithmic scale) of various types of M-D-M filters. The "double filter" consists of two identical filters cemented together to form a composite filter. Although the peak transmittance is low, (T_{max} is from 0.10 to 0.30), an extremely high attenuation in the reject region is attained - the optical density is greater than five. This high attenuation is achieved because the filters are absorbing and hence the considerations which apply to dielectric films (see 20.5.1.4) do not hold here.

20.10.3.2 The wedge M-D-M filter. Another interesting form of a M-D-M filter is the wedge Fabry-Perot type filter,[66] which is depicted in Figure 20.107. Two silver films are deposited on either side of a layer of magnesium fluoride, which is wedged shaped so that its thickness varies in a linear fashion along the length of the filter. At one end the optical thickness is a half-wave of violet light (400 mμ), so this portion passes the violet. At other positions along the filter the dielectric film is thicker and these portions pass the blue, green, yellow, and finally the red.* In actual practice, the filter is usually manufactured a second order filter, rather than first order, as shown, thus achieving a narrower band width. In this case, at the thick portion of the wedge the second order red and the third order blue overlap, and it is necessary to remove the blue by appropriate dye or glass filters. The $\Delta\lambda_{1/2}$ of this second order filter is 10 mμ, independent of wavelength. The slope of the wedge and the length of the filter are chosen so that a one millimeter slit gives this pass band. Thus by inserting a slit 1 mm wide in front of this wedge filter and illuminating it with white light, a rather inexpensive source of quasi-monochromatic light is obtained. This type of wedge can be deposited in an annular ring on a disk.[67] The wavelength scanning is accomplished by rotating the disk past a slit.

20.10.3.3 Other types of narrow pass-band filters which contain metal films. A higher peak transmission and narrower band width is attained with F-P type filters if several dielectric films are added to the stack in addition to the spacer layer.[64] Turner and Berning have devised some band-pass filters which contain a single silver film and many dielectric films.[68] M-D-M filters are also useful as reflection filters, particularly in the infrared spectral region.[64, 69]

20.10.4 All-dielectric Fabry-Perot type filters. The simplest form of this type of filter is shown in Figure 20.108. The silver films are replaced by semitransparent mirrors composed of dielectric materials, such as a quarter-wave stack. Thus the design of a filter consisting of seven layers would be:

glass H L H LL H L H air .

Here the H L H combination is a three-layer quarter-wave stack (H and L have a QWOT at λ_o) and LL represents a spacer layer of half-wave optical thickness at λ_o. The spectral transmission versus frequency of such a multilayer is shown in Figure 20.109. Although this and other multilayers which are used as illustrations can be used only in the infrared because they contain germanium as a high index material, the principles involved here apply to any spectral region.

20.10.4.1 T_{max}. The concept of an absentee layer (20.1.5.2.2) is useful in determining the transmission of this filter at the wavelength λ_o (i.e. $g = 1.0$) where the maximum of the pass band is located. At this wavelength the LL layer is absentee and hence it can be removed from the stack, leaving:

glass H L H H L H air.

This leaves two of the H layers next to each other, resulting in the layer HH which has an optical thickness of a half-wave. After removing this HH combination from the stack we are left with four layers:

glass H L L H air.

---

* In actual practice, the optical thickness of the dielectric spacer layer is slightly thinner than a half-wave, due to the phase shift upon reflection of the silver films. Also, the thickness of silver films varies along the wedge, due to the dispersion of the optical constants of the silver.

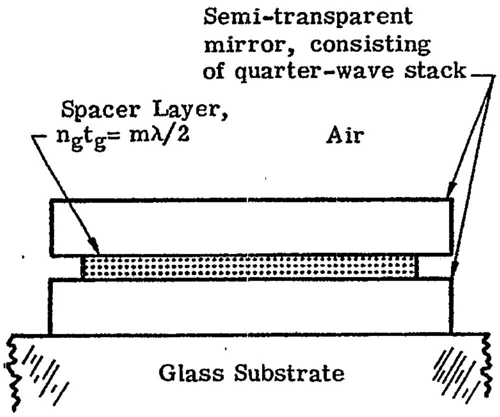

Semi-transparent mirror, consisting of quarter-wave stack
Spacer Layer, $n_g t_g = m\lambda/2$
Air
Glass Substrate

Figure 20.108- Design of a conventional all-dielectric Fabry-Perot type filter.

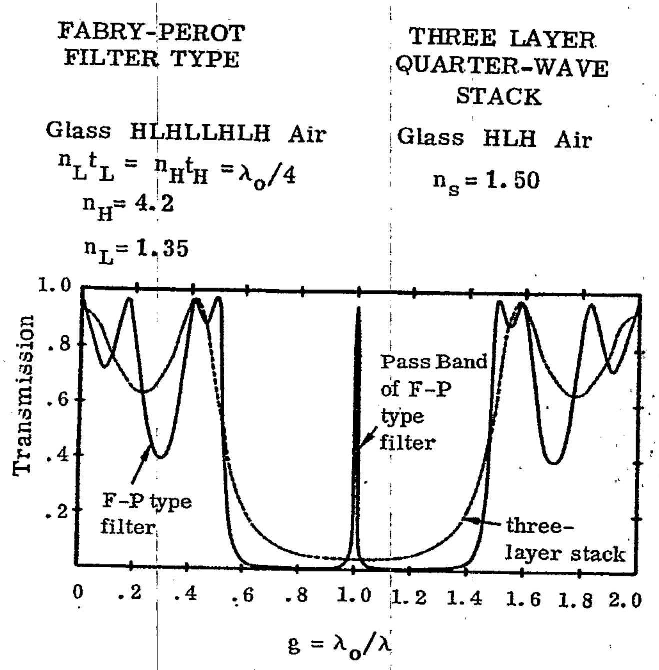


Figure 20.109- Computed spectral transmission of an all-dielectric Fabry-Perot type filter (solid curve) and a three-layer quarter-wave stack (dashed line).

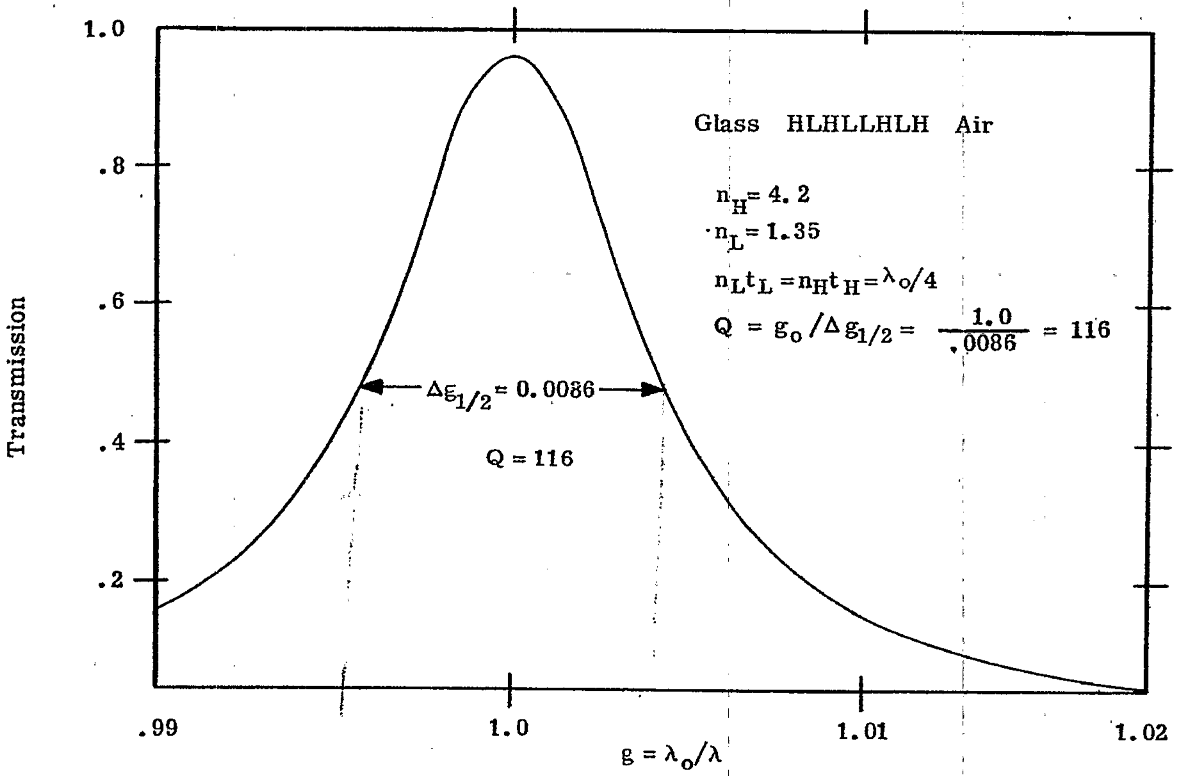


Figure 20.110- Computed transmission of the pass band of the Fabry-Perot type filter shown in Fig. 109.

The half-wave combination LL is absentee and can also be removed from the stack. Repeating this process, we see that at λ_o the transmission is the same as that of a single surface of uncoated glass of index 1.50, and thus from Equation (21), $T = 0.96$. This does not include the reflection loss at the second surface of the substrate.

20.10.4.2 It is instructive to compute the Q of this filter. First, it is necessary to compute the reflectivity, R_1 and R_2, of each of the effective interfaces. The first effective interface is:

glass H L H cryolite

and thus R_1 is computed from an incident medium of cryolite. R_2 is computed for a stack:

air H L H cryolite

where the substrate is air and cryolite $(n_o = 1.35)$ is the incident medium. Utilizing Equations (40) and (43), we find that $R_1 = 0.953$ and $R_2 = 0.968$, whence $R = (R_1 R_2)^{1/2} = 0.961$. From Equation (58) a Q of 80 is computed. Figure 20.110 shows the computed spectral line shape of this filter, on a frequency scale. The Q, measured from this graph, is 116. The additional narrowing of the transmission band is attributed to the phase shift upon reflection of the three layers which constitute the effective interfaces of this F-P type filter. References 56 and 61 show how to include the effect of the phase shift in computing the Q of the system.

20.10.4.3 Effect of phase shift upon reflection. As an example of how the phase shift upon reflection can influence the shape of the transmission band of a F-P type filter, consider the following multilayers, which are designated as Design I and Design II :

I glass L H L HH L H L air

II glass HH L H L HH L H L HH air .

Design II is essentially Design I with an extra half-wave layer added to each end of the stack. At $g = 1.0$, the half-wave layers are absentee and the reflectivities of the effective interfaces of each of the two stacks is exactly the same. However, if we examine the transmission bands for these two stacks, shown in Figure 20.111, it is evident that the width $\Delta \lambda_{1/2}$ at $1/2\ T_{max}$ is somewhat less for Design I than for Design II. This can be attributed to the variation with wavelength phase shift upon reflection of the effective interfaces, which is shown in Figure 20.112. In each case the phase shift upon reflection is measured from inside of the germanium spacer layer. At $g = 1.0$ the phase shift is zero, which means that the node lies at the surface of the multilayer. At lower frequencies than $g = 1.0$ the node lies to the right of the surface. The shift of this node as the wavelength changes alters the shape of the transmission band.

20.10.4.4 The use of blocking filters. All-dielectric F-P type filters are usually used in conjunction with blocking filters. These blocking filters must in general have a much higher attenuation than the blocking filters which are used in conjunction with the M-D-M type filters. The reason is that in the former case, the unwanted transmission bands cover a wide spectral region, whereas the unwanted transmission bands of a M-D-M type are usually quite narrow. For example, suppose it is desired to use the transmission band at 477 mμ in the M-D-M type second order filter shown in Figure 20.105. In this case it is necessary to use an auxiliary blocking filter to eliminate the unwanted transmission bands below 425 mμ and the band at 712 mμ . There are many absorption type filters which could be used to attenuate below 425 mμ. The band at 712 mμ is comparatively narrow and hence the total amount of radiant flux which "leaks" through this band is not large. Consequently, the amount of attenuation required in the blocking filter is not as great as it would be if the pass band were wide. In the case of all-dielectric F-P type filters, the quarter-wave stacks which are used for the semitransparent mirrors do not reflect over a wide range of wavelengths. Thus a quite appreciable amount of radiant energy is liable to "leak" through the filter in the region outside of the high-reflectance zone of the mirrors. The narrower the pass band of the F-P type filter, the more effective should be the blocking filters to insure that the total amount of radiant flux transmitted in the spectral region of the pass band of the filter should be much greater than the flux which leaks through at other wavelengths. For example, the all-dielectric F-P type filters shown in Figures 20.115 and 20.116 have a substantial transmittance below 8.0 μ. The total transmitted radiant flux below 8.0 μ to 3.9 μ (where the PbTe films start to absorb) is considerably larger than the radiant flux transmitted through the pass band near 10 μ.

20.10.5 All-dielectric F-P type filters for the visible. All-dielectric F-P type filters are produced commercially for most of the visible spectral region with a wide range of Q . Filters with a Q from 10 to 1000 are available. Values of T_{max} from 0.45 to 0.60 are commonly attained, which includes the appropriate blocking filters. The spectral transmission of an all-dielectric F-P type filter which has its pass

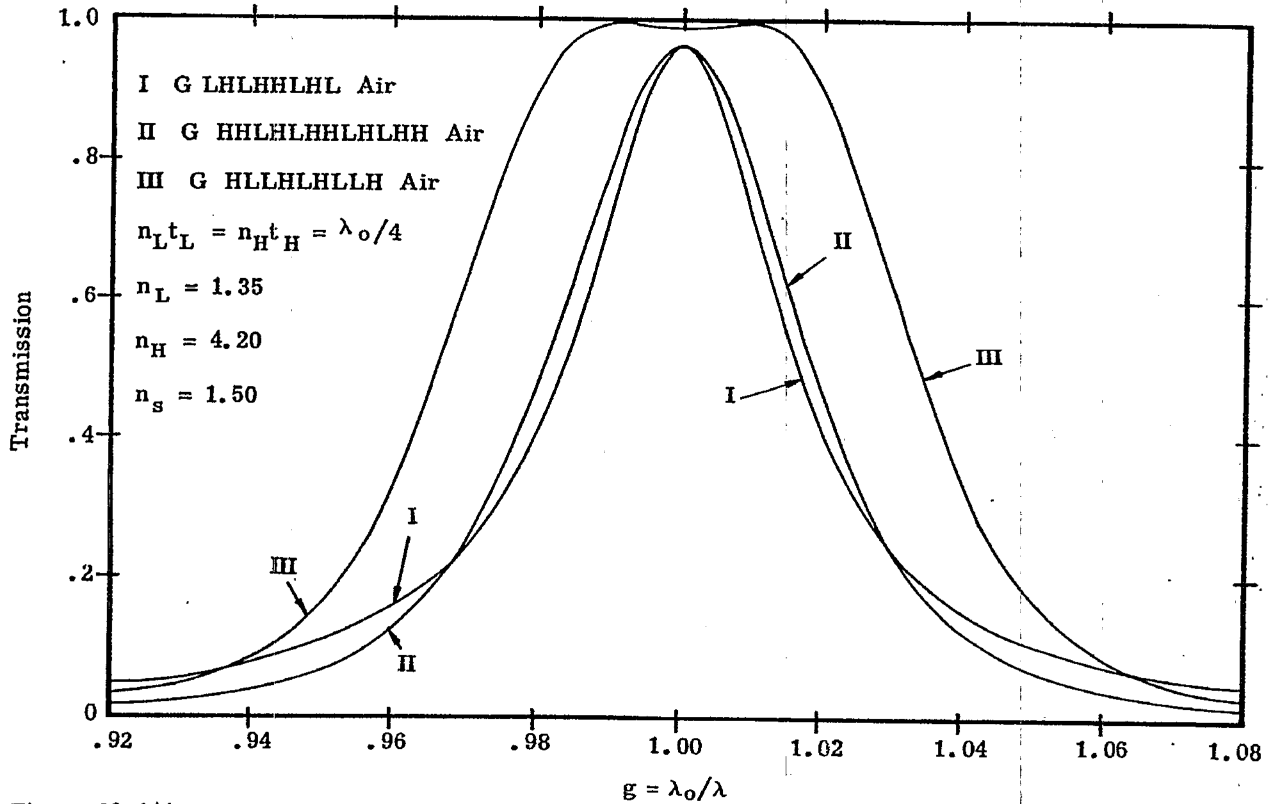


Figure 20.111- Computed spectral transmission of the pass bands of some all-dielectric Fabry-Perot type filters. Filter III has non-Lorentzian shaped pass band.

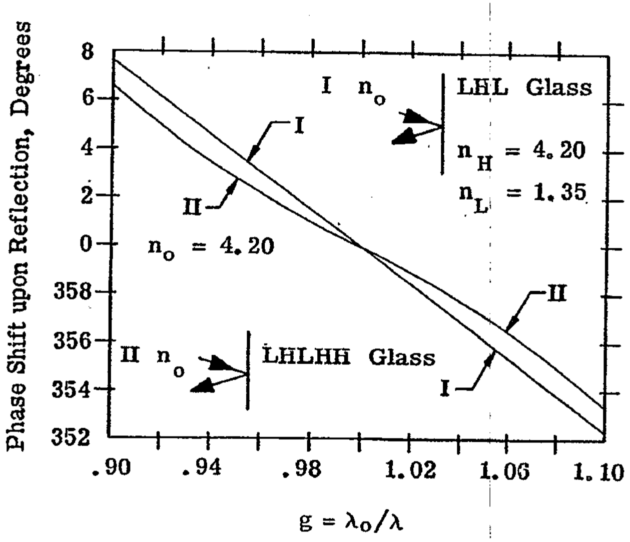


Figure 20.112- The phase shift upon reflection, ϵ_s of the "effective interface" of the filters I and II in Fig. 111.

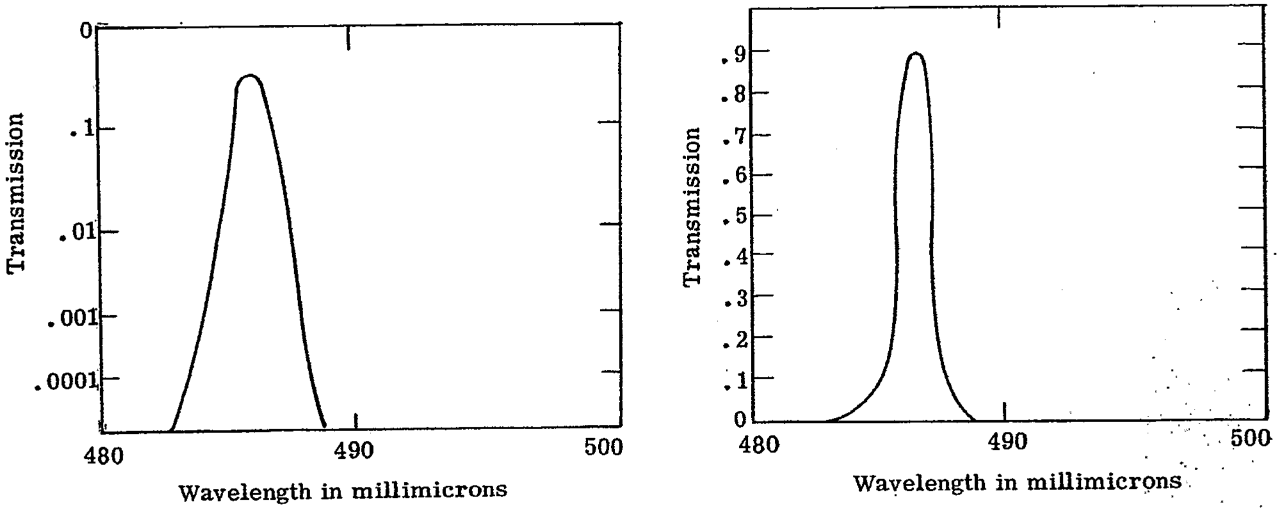


Figure 20.113- Measured transmittance of an all-dielectric Fabry-Perot type filter on a linear and logarithmic scale. Courtesy of Spectrolab, a Division of Textron Electronics, Inc.

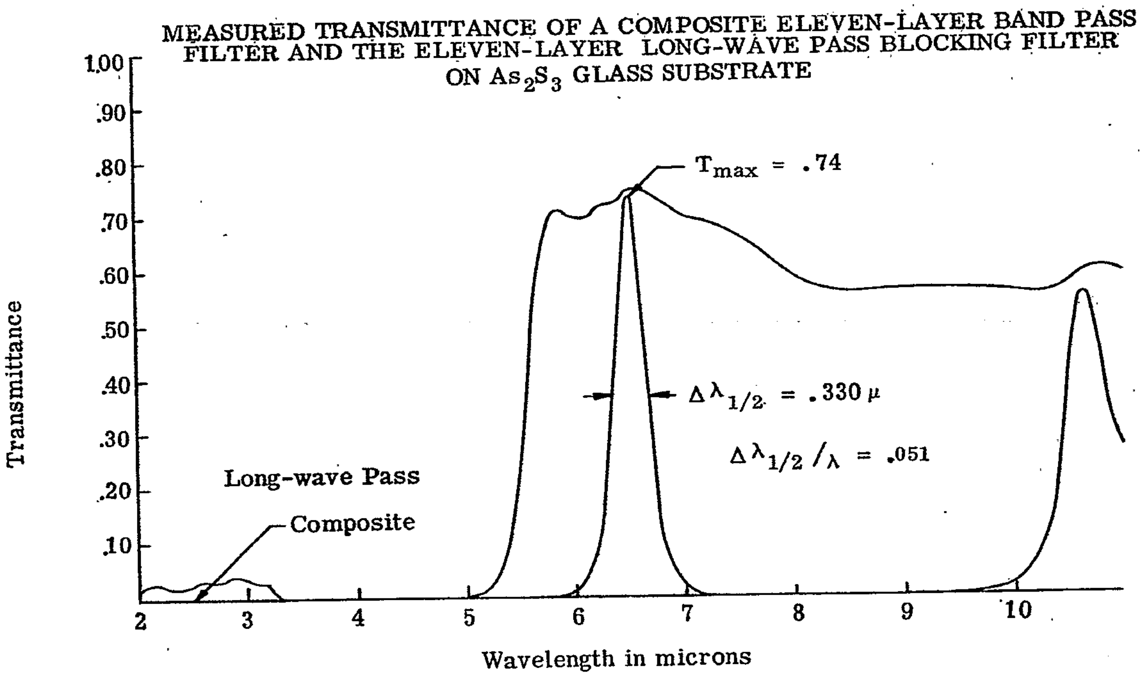


Figure 20.114- Measured spectral transmittance of a Fabry-Perot type filter and auxillary blocking filter composed of ZnS and PbTe .

band in the blue, is shown in Figure 11- . Although the peak transmission would be lowered slightly if a blocking filter were added in tandem, the T_{max} is still quite high. Also, the attenuation is quite high; an optical density of 4.0 is achieved at $3 \Delta \lambda_{1/2}$ from the wavelength of maximum transmission.

20.10.5.1 Extremely narrow bandwidth filters. From theoretical considerations alone, one might conclude from Equation (58) that it is possible to construct all-dielectric multilayer F-P filters which have Q's of 10,000 and hence a band width of 0.05 mμ at 500 mμ in the visible. Such filters would be quite useful to astronomers, who have been using the expensive Lyot type polarization filters to isolate the H^{α} line. These filters would also find many applications in spectrochemical analysis; in some instances they could replace costly spectrometers which contain diffraction gratings. One method of attaining such large values of Q would be to increase the number of layers in the quarter-wave stacks which constitute the effective interfaces, and thus obtain higher reflectivities. The same practical difficulties which were described in 20.8.2.3 are encountered. The small amount of absorption and scattering in each film degrades the reflectivity and consequently sets a lower limit on the bandwidth of F-P type filters. Thus it is not too difficult to manufacture all-dielectric F-P type filters for the visible spectral region with a $\Delta \lambda_{1/2}$ of from 1.0 mμ to 1.5 mμ and a T_{max} of greater than 0.50 (this does not include blocking filters). A Russian publication[65] reports that filters with a $\Delta \lambda_{1/2}$ of 0.13 mμ and a T_{max} of 0.15 have been produced.

20.10.5.2 Filters which use a mica spacer. Another method of increasing the Q of a filter is to increase the order of interference m . The mechanical stress in the films (Section 20.2.3.2.4) makes it impractical to use thick layers of evaporated material to manufacture a spacer layer with a high order of interference. By using a thin sheet of mica as a spacer in a F-P type filter,[70, 71] it is possible to attain values of m in the range from 70 to 700. Both sides of the mica spacer, which is usually from 0.005 to 0.0005 inches in thickness are coated with a semitransparent multilayer mirror, such as a quarter-wave stack. Using this technique, a filter which isolates one of the lines of the yellow sodium doublet has been produced.[71] A filter with a pass band at 570mμ , a $\Delta\lambda_{1/2}$ of 0.1 mμ and a T_{max} of 0.25 is reported.[71] The difficulty of using filters with a high order of interference is mentioned in 20.10.1.2, namely the problem of blocking out adjacent transmission bands which, in the case of the filter cited in the foregoing sentence, occur at intervals of 1.1 mμ on either side of the main pass band. This can be accomplished by inserting additional mica F-P type filters in tandem, but this reduces the T_{max} .

20.10.5.3 F-P filters at non-normal incidence. If a F-P type filter of the type shown in Figure20.108or20.103 is inserted in a collimated beam of light at non-normal incidence, the following effects are observed as ϕ increases:

(1) The transmission pass band broadens and shifts to shorter wavelengths (i.e. a blue shift). This is because $n_g t_g$ in Equation (55) is replaced by an effective thickness (see 20.1.6.2) which is less than its original value. Hence a smaller λ_o satisfies this equation.

(2) The transmission band is partially linearly polarized. If the incidence angle ϕ is increased to large enough angles, two distinct bands are seen, each at a different wavelength. The light in one band is linearly polarized in the "s" plane, and the other in the "p" plane.

This angle shift of the maximum of the pass band can often be used to good advantage. For example, suppose it is desired to isolate the mercury green line at 546μ, and a F-P type filter which is available has a pass band at 550mμ. The spectral position of the pass band can be easily shifted so that it passes the Hg green line by tilting the filter less than ten degrees. Of course, the performance of the filter has been degraded by this tipping because the pass band has been broadened, but the loss is not serious if the pass band is wide to begin with. It is also evident that if a F-P type filter is placed in a beam of convergent light, then the angle shift broadens the transmission band asymmetically towards shorter wave-lengths. Thus a filter which is placed in a convergent beam should have its λ_o at normal incidence at a slightly longer wave-length. For example, Lissberger and Wilcock,[72a, 72b] calculate that a filter which is to have its optimum performance at 5000 mμ is placed in an f 2.0 beam, then at normal incidence its T_{max} should be located at $\lambda_o = 502$ mμ. It is also evident that filters which have extremely narrow band widths should be used at normal incidence, in a well collimated beam to prevent the loss of the narrow bandwidth by the angle shift broadening. The spectral transmittance curve in Figure 20.115 shows the angle shift of an all-dielectric F-P type filter for the infrared.

20.10.6 All-dielectric F-P type filters for the infrared. All-dielectric F-P filters are available for the infrared spectral region with a Q as large as 200. Figures 20.114, 20.115, and 20.116 show the spectral transmittance of some all-dielectric F-P type filters which are used in the infrared. These curves are intended to present a sample of what can be accomplished. The filter shown in Figure 20.114 is intended to have a

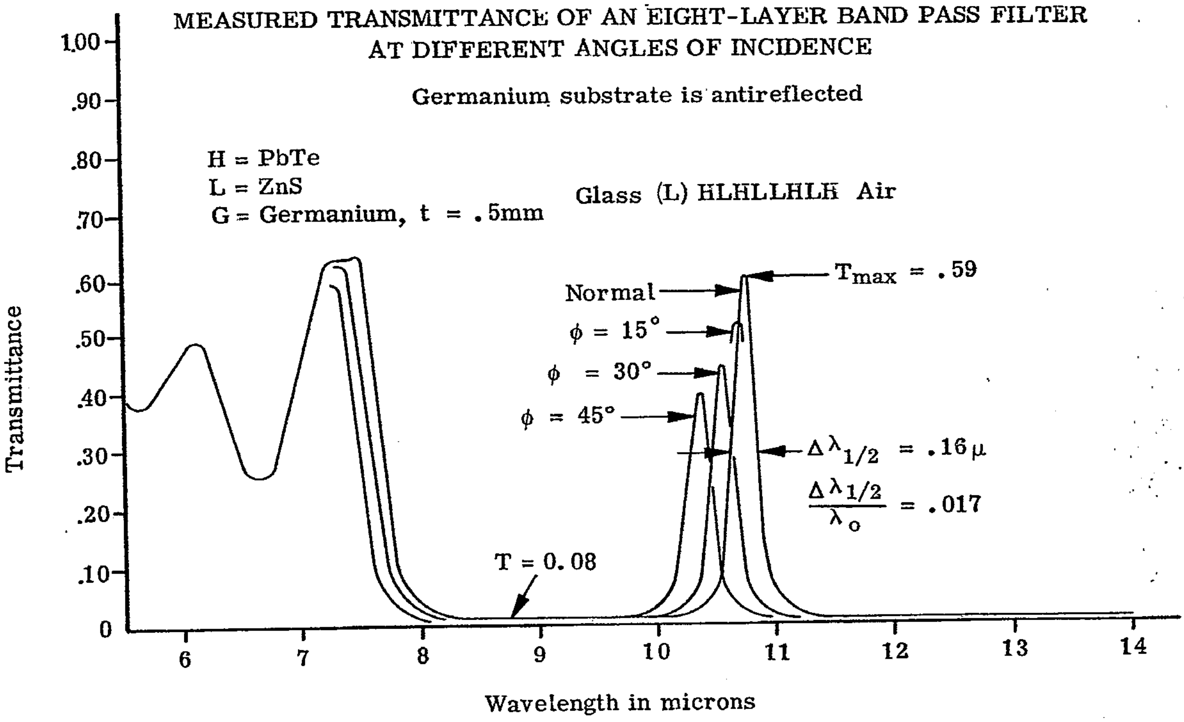


Figure 20.115- Measured spectral transmittance of a Fabry-Perot type filter at various values of ϕ. The decrease in T at $\phi = 45°$ is due to vignetting in the spectrophotometer. Courtesy of Bausch and Lomb, Inc.

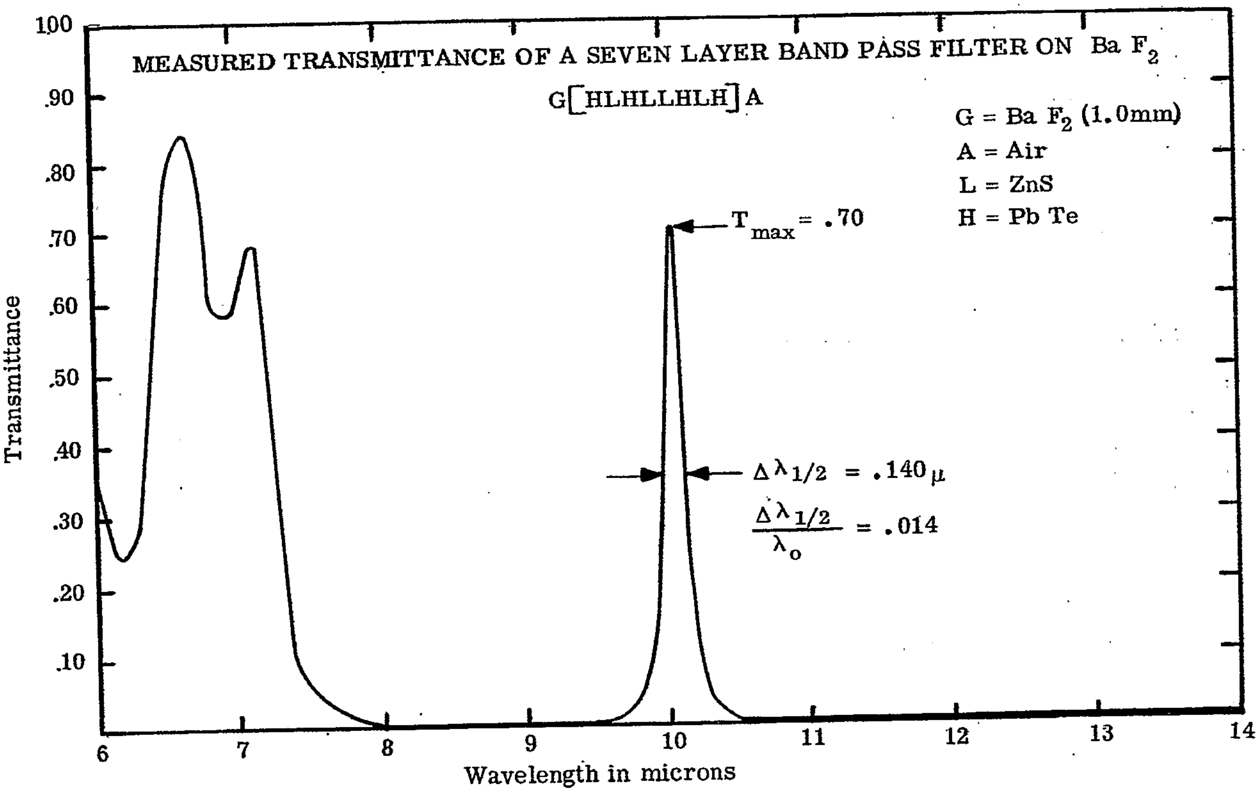


Figure 20.116- Measured spectral transmittance of a Fabry-Perot type filter on a BaF_2 substrate. Courtesy of Bausch and Lomb, Inc.

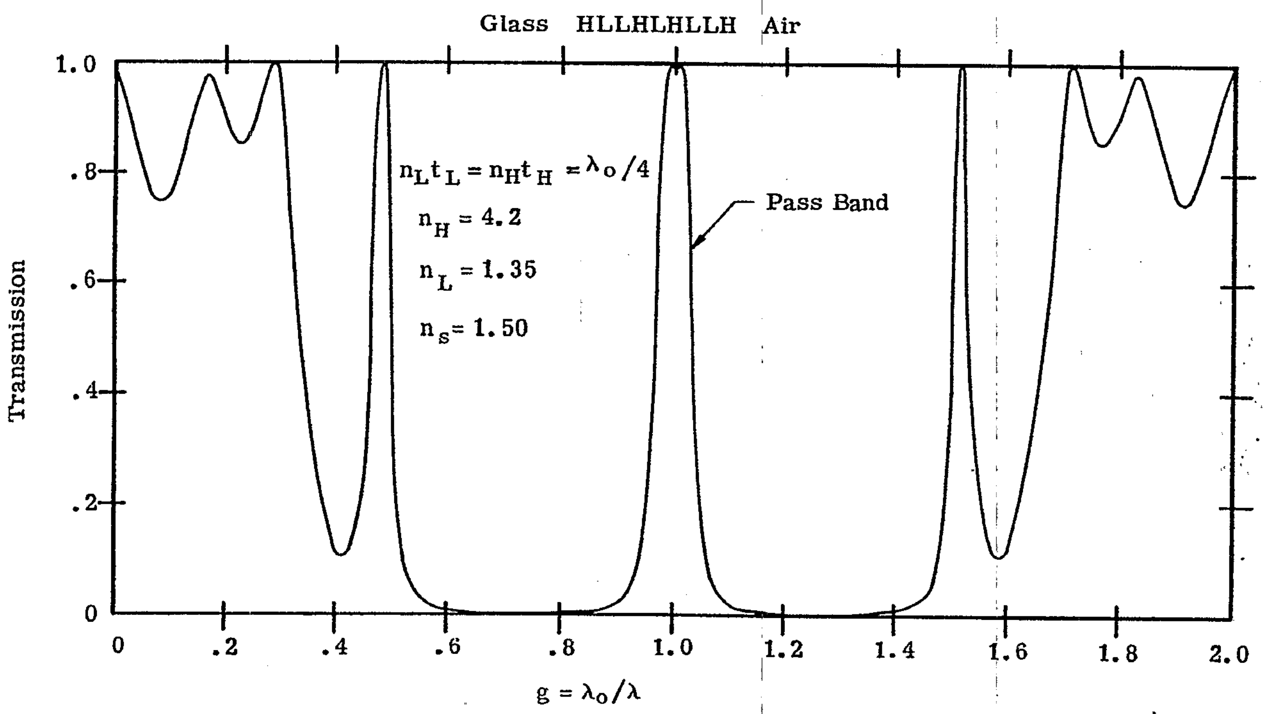


Figure 20.117- Computed spectral transmission of a Fabry-Perot type filter which has a non-Lorentian shaped pass band.

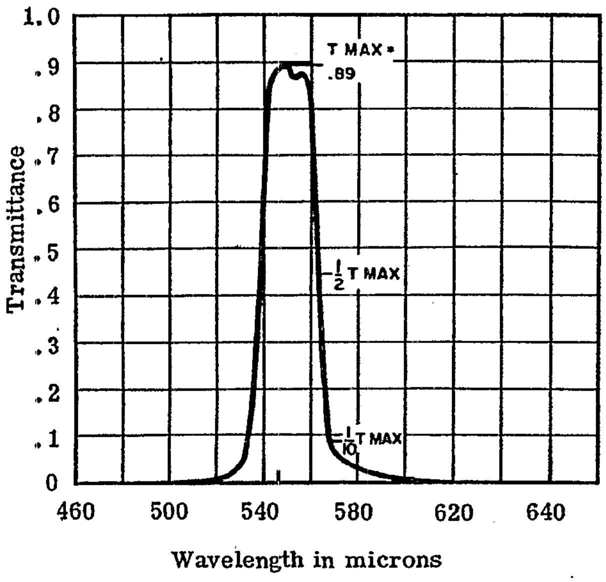


Figure 20.113 - Measured spectral transmittance of a narrow band-pass filter which has a nearly rectangular shaped pass band. Courtesy of Baird-Atomic, Inc.

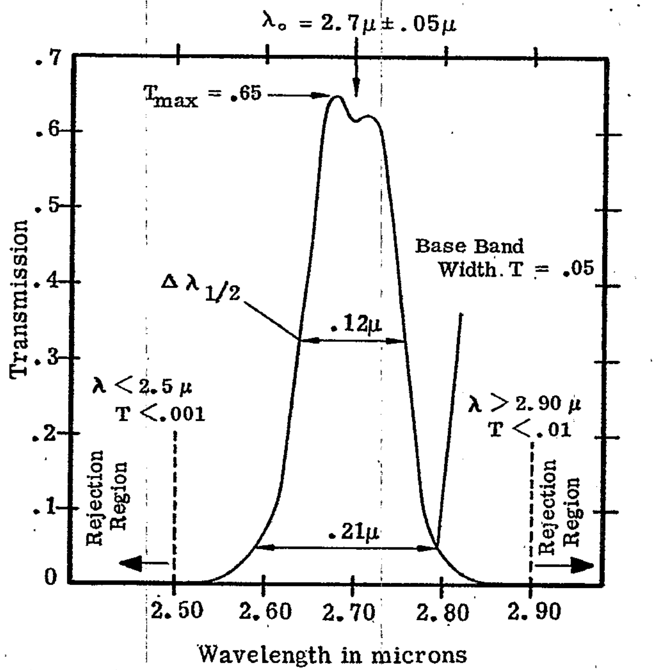


Figure 20.119- Measured spectral transmittance of a narrow band-pass filter which has a nearly rectangular shaped pass band. Courtesy of Eastman Kodak Company.

broad pass band; the Q is about 20. A long-wave pass blocking filter is inserted in tandem to attenuate the unwanted transmission "leak" below 3.3 μ. The region of high transmission above 10 μ is outside of the high-reflectance zone of the dielectric mirrors in the effective interfaces. The F-P type multilayer shown in Figure 20.115 is deposited on a substrate of germanium. The "L" layers which are in parenthesis, sic. (L), are antireflection coatings for the germanium substrate. At λ_o, which is close to 10.8 μ at normal incidence, the same analysis which was applied in 20.10.5 is used here to remove absentee layers from the stack, thus leaving (L) germanium (L), a germanium slab with an "L" antireflection coating on either side. Although the refractive index of this L layer (see 20.3.3) is not the optimum value, it still improves considerably transmission at λ_o. If these (L) coatings were not added, then T_{max} would be less than 0.47, instead of the value of 0.59 which is shown. The angle shift of the pass band to shorter wavelengths is also shown in Figure 20.115. The decrease in the T_{max} at non-normal incidence can be attributed to vignetting in the spectrophotometer which measured the transmittance. Figure 20.116 shows essentially the same type of coating, but on a substrate of barium fluoride. The substrate is a low-index material, and hence it is unnecessary to add the (L) coatings to antireflect the substrate, as is necessary for the multilayer shown in Figure 20.115. Comparing Figures 20.115 and 20.116, the effects of using a different substrate are manifested in the greater T_{max} for the multilayer on the CaF_2 substrate, a slightly narrower width of the pass band, and the higher transmission in the short-wave region below eight microns. Of course the latter effect is undesirable in many applications; the "leak" in the transmission in the short-wave region of the multilayers shown in Figures 20.115 and 20.116 could be removed by the addition of suitable blocking filters, as is done in Figure 20.114. Lead telluride is used as a high-index film material in this spectral region because it is transparent and has a large refractive index - and hence a filter with quite respectable Q is obtained with a small number of layers. Zinc sulfide is used as the low index material because it does not have a large mechanical stress (see 20.2.4.2.4) and also because it is transparent in this long-wave region. Silicon monoxide is not used in this spectral region because it has a strong absorption band starting at 8 μ.[72c] Greenler[16] has fabricated Fabry-Perot type filters with pass bands in the 10 μ region by using tellurium as a high-index material and sodium chloride as a low-index material. Using the same materials it is possible to manufacture filters with a pass band at wavelengths as long as 20 μ.

20.10.7 F-P type filters with a pass band of non-Lorentzian shape. The F-P type filters which are described in 20.10.4 and 20.10.5 have a transmission pass band which is essentially Lorentzian in shape. The main drawback of this type of line shape is mentioned in 20.10.1.6, namely that unless the pass band is narrow, the filter has a long transmission tail which decreases in amplitude quite slowly. For example, suppose it is desired to use a multilayer filter to isolate the emission line at 491.6 mμ of a mercury discharge lamp. The lamp does not emit lines of any strength for at least 40 mμ on both the short-wave and long-wave side of 491.6 mμ. Therefore, it is not important that the filter have a narrow transmission band, in fact, the $\Delta\lambda_{1/2}$ could easily be as large as 20 mμ, provided the discharge tube does emit an appreciable amount of continuum radiation. However, it is quite important that the emission lines at 435.8 mμ in the blue and at 546.1 mμ in the green be attenuated very effectively, because these lines are at least a thousand times more intense than the 491.6 line. Thus, if the 491.6 mμ which passes through the filter is to be merely ten times more intense than the light from the blue and green lines, then the transmission of the filter is at 496 mμ and 546 mμ must be at least 10^{-4} of T_{max}. A M-D-M type filter would not furnish this much attenuation. A double filter of this type would furnish this much attenuation, but at the expense of a very low T_{max}. An all-dielectric F-P multilayer would furnish this degree of attenuation, provided that a large number of layers were used in the filter. However, this would mean that the transmission band would be quite narrow, whereas a narrow transmission band is not requisite. Such a filter with a narrow band would be expensive for two reasons: First, it contains a large number of layers, and is expensive to manufacture. Second, because the transmission band is narrow, the thickness of the layers must be controlled to quite close tolerances, so that the peak transmission of the filter occurs at exactly the desired wavelength of 491.6 mμ. The latter difficulty is avoided if a multilayer has a pass band which is essentially rectangular in shape, as is shown in Figures 20.117 to 20.121. Space does not permit us to elaborate some of the methods which are used to achieve transmission bands of this shape. One filter of this type is called a double half-wave system; the theory of such filters is discussed by Smith.[60] The spectral transmission of such a filter is shown in Figure 20.117. When analyzed as a F-P type filter, each of the effective interfaces is the film combination H LL H and contains a half-wave film; hence the name, double half-wave. The spacer layer is a quarter-wave optical thickness, rather than a half-wave layer. Figure 20.111 shows the transmission in the pass region so that it can be compared with the Lorentzian-shaped transmission bands of a conventional all-dielectric filter shown in the same Figure. Figure 20.118 shows the transmittance of a multilayer of this type with its pass band in the visible spectral region, while Figures 20.119, 20.120, and 20.121 depict infrared filters which pass at 2.70 μ, 4.50 μ, and 10.8 μ, respectively. Additional blocking filters have been added to the filters whose transmission curves are shown in Figures 20.119 and 20.120.

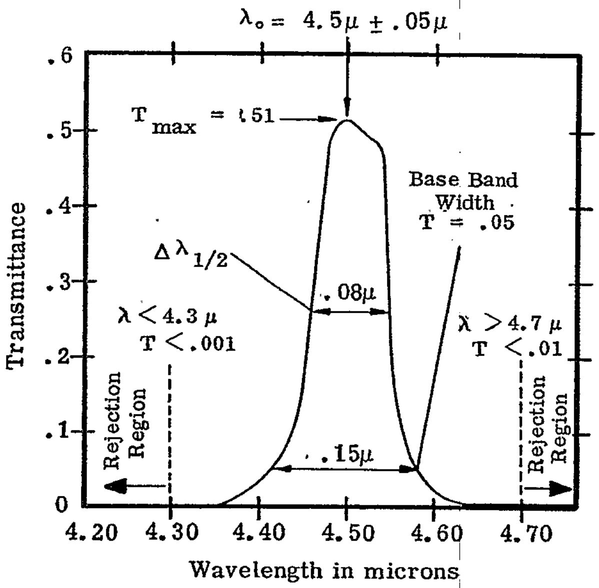


Figure 20.120- Measured spectral transmittance of a narrow band-pass filter which has a nearly rectangular shaped pass band. Courtesy of Eastman Kodak Company.

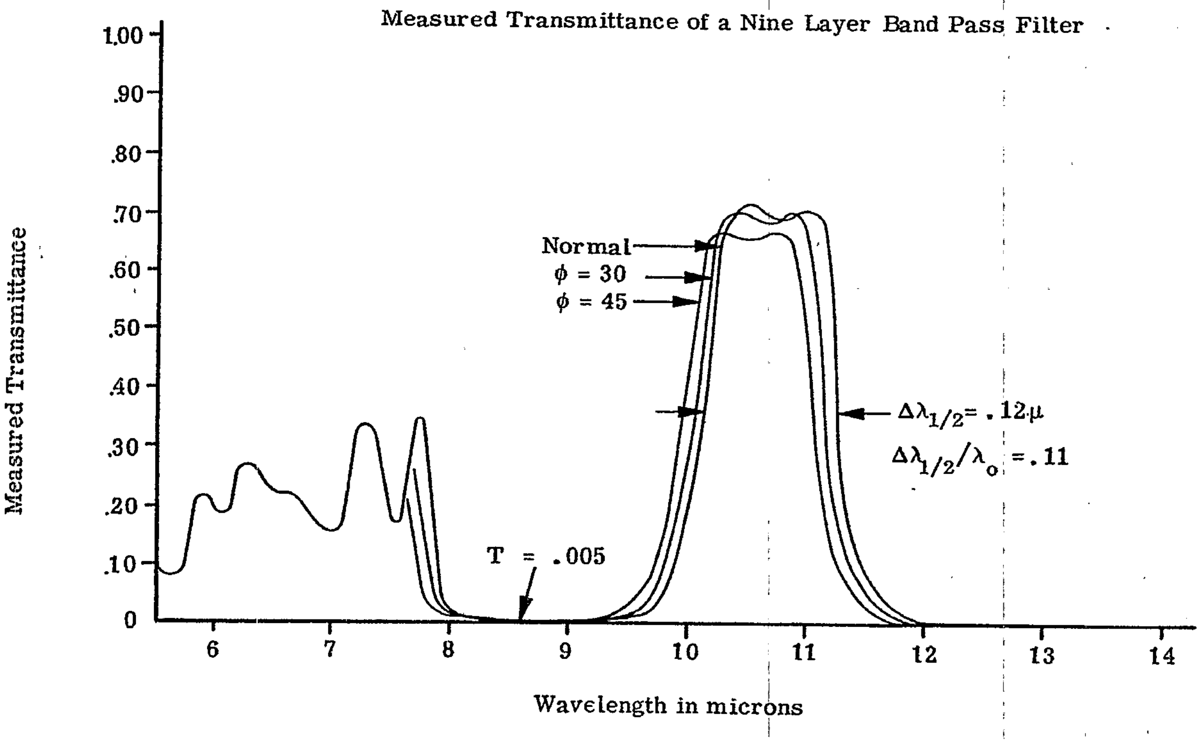


Figure 20.121- Measured spectral transmittance of a narrow band-pass filter which has a nearly rectangular shaped pass band. Courtesy of Bausch and Lomb, Inc.

20.11 REFERENCES FOR FURTHER STUDY

Publications on thin film optics can usually be placed in either of two categories:

(1) Thin film science: This is a study of the fundamental properties of thin films, with an emphasis on knowledge for the sake of knowledge, rather than on knowledge for the sake of ultimately producing some gadget. This includes a study of the physical structure and optical properties of thin films, the thermodynamics and chemistry of their formation, the properties of the films as related to the physics of solids, and so on.

(2) Thin film technology. Section 20 has been devoted to one aspect of this topic, namely how our knowledge of thin films can be utilized to provide useful optical components, such as filters, beam splitters, and other devices. Included in this broad classification are methods and techniques of preparing films, controlling their thickness, methods of multilayer filter design, and so on. The optical applications of thin film technology began to expand rapidly after 1946. A conference on thin film optics was held in 1950 and the papers which were presented at this conference[73] are a good summary of the state of the art up to that time. In 1955 Heavens[7] published a book which not only covers many aspects of thin film science, but he also devotes the later part of this book to some applications of thin films. Therein is presented some of the topics which have not been covered in Section 20, such as multilayer polarizers, frustrated total reflection filters, and so on. Another publication of Heavens[21] surveys more recent developments. Both Weinstein (Welford)[5] and Abeles[12] both present terst and correct mathematical treatments of the theory of the reflectivity of multilayers. Vasicek's book[74] contains some useful information, but very often it is hidden in endless pages of redundant and repetitive derivations of recursion formulae. The practical aspects of depositing thin film coatings are lucidly presented in a Navy Department pamphlet.[75] Holland's book[18] is more recent and presents a vast amount of valuable lore on how to evaporate thin films in a vacuum.

REFERENCES

1. Heavens, O. Optical Properties of Thin Solid Films (Butterworth Scientific Publications, 1955) p. 238
2. ibid, p. 221
3. Strong, John. Concepts of Optics (Freeman, 1958) p. 251
4. Heavens, O. op cit p. 66
5. Weinstein, W. (W. Welford) Vacuum 4, 3 (1954)
6. Terman, F. E. Radio Engineering (McGraw Hill, 1947) 3rd. ed. p. 75
7. Drumheller, Carl "Silicon Monoxide Evaporation Techniques" - A monograph available from the Kemet Company, Cleveland, Ohio. (1960)
8. Drumheller, Carl "Properties and Application of Silicon Monoxide" - A monograph available from the Kemet Company, Cleveland, Ohio. (1960)
9. Weinstein, W. op cit p. 5
10. Heavens, O. op cit Chapter 4
11. Born, M and Wolf, E. - Principles of Optics (Pergamon, 1959) p. 54
12. Abelès, F. Ann. de physique, 5, 103 (1950)
13. Welford, op cit p. 5
14. Pohlack, Hubert "Zum Problem der Reflexionsminderung optisher Gläser bei nichtsenkrechtem Lichteinfall" Jenaer Jahrbuch (VEB Optik Carl Zeiss Jena, 1952)
15. Hass, G. and Tousey, R. J. Opt. Soc. Am. 49, 593 (1959)
16. Greenler, R. G. J. Opt. Soc. Am. 47, 130 (1957)
17. Hass, George and Turner, A. F. "Preparation of Thin Films" - in Volume 6 of Methods of Experimental Physics, L. Marton, Editor in Chief. (Academic Press, 1959)
18. Holland, L. - Vacuum Deposition of Thin Films (Chapman and Hall, 1956)
19. Ballard, S., McCarthy, K. A. and Wolfe, W. L. State-of-the-Art Report: Optical Materials for Infrared Instrumentation. (Report No. 2389-11-S: I.R.I.A, Univ. of Michigan, 1959)
20. Turner, A. F. et al - Thick Thin Films - Quarterly Technical report #4 under contract with U.S. Army Engineer Research and Development Laboratories, Fort Belvoir, Va. (1951)
21. Turner, A. F. and Truby, F. K. U. S. Patent 2,858, 240(Issued October 1958)

22a Heavens, O. Reports on Progress in Physics, 23,60 (1960)

22b Hass, G., Ramsey, J. B. and Thun, R. J. Opt. Soc. Am. 49, 116 (1959)

23. Huldt, Lennart and Staflin, T. Optica Acta 6, 27 (1959)
24. Hass, G. Vacuum 2, 331 (1952)
25. Pohlack, Hubert. loc cit p. 106
26. Baumeister, P. Optica Acta 8, 105 (1961)

REFERENCES (continued)

27. Hass, George and Turner, A. F. "Coatings for Infrared Optics" in Ergebnisse der Hochvakuumtechnik und der Physik dünner Schichten (Wissenschaftliche Verlagsgesellschaft Stuttgart, 1957)

28. Cox, J. T., Hass, G. and Jacobus, G. F., J. Opt. Soc. Am. 51, 714 (1961)

29. Cox, J. T., Hass, G., and Rowntree, R. F. Vacuum 4, 445 (1954)

30. Turner, A. F., Berning, P. et. al. Infrared Transmission Filters, Quarterly Technical Report #6, under contract DA44-009-eng-1113 with the U. S. Army Engineer Research and Development Laboratories, Fort Belvoir, Va. (1953)

31. Hass, G. and Cox, J. T. Published in Vol. 52 of J. Opt. Soc. Am. (1962)

32. Turner, A. F., Epstein, I., et. al. Optical Properties of Multilayer Films and Interference Filters in the 10 Micron Region. Quarterly Technical report #5 under contract with the U. S. Army Engineer Research and Development Laboratories, Fort Belvoir, Va. (1951)

33. Brillouin, Leon. Wave Propagation in Periodic Structures. (McGraw-Hill, 1946 or Dover, 1953)

34. Seitz, Frederick, The Modern Theory of Solids Chapter 9 (McGraw-Hill, 1940)

35. Epstein, I. J. Opt. Soc. Am. 42, 806 (1952)

36. Baumeister, P. J. Opt. Soc. Am. 48, 955 (1958)

37. Dimmick, G. L. and Widdop, W. E. J. Soc. Motion Picture Engrs. 58, 36 (1952)

38. Carlson, F. E., Howard, G. T., Turner, A. F. and Schroeder, H. H. J. Soc. Motion Picture Engrs. 65, 136 (1956)

39a Turner, A. F. & Schroeder, H. H. J. Soc. Motion Picture Engrs. 69, 351 (1960)

39b Koch, George U. S. Patents 2,552, 184 and 2,552, 185

40. Epstein, L. I. J. Opt. Soc. Am. 45, 360 (1955)

41. Thelen, A. "The Use of Vacuum Deposited Coatings to Improve the Conversion Efficiency of Silicon Solar Cells in Space." Progress in Astronautics and Rocketry Academic Press, (1961) p. 373

42. Sennett, R. S. and Scott, G. D. J. Opt. Soc. Am. 40, 203 (1950)

43. Holland, L. Vacuum 3, 159 (1953)

44. Turner, A. F. and Schroeder, H. H. J. Soc. Motion Picture Engrs. 61, 628 (1953)

45. American Institute of Physics Handbook (McGraw-Hill, 1957) p. 6-108

46. Jenkins, F. A. J. phys. radium 19, 301 (1958)

47. Kuhn, H. and Wilson, B. A. Proc. Phys. Soc. (London) B 63, 754 (1950)

48. Stone, J. M. J. Opt. Soc. Am. 43, 927 (1953)

49. Baumeister, P. W. and Stone, J. M. J. Opt. Soc. Am. 46, 228 (1956)

50. Ring, J. and Wilcock, W. L. Nature 171, 648 (1953)

51. Giacomo, P. J. phys. radium 19, 307 (1958)

52. Giacomo, P. Rev. Opt. 35, 317 (1956)
35, 442 (1956)

REFERENCES (continued)

53. Jenkins, F. A. and White, H. E. Fundamentals of Optics (McGraw-Hill, 1957) Third Ed. p. 273

54. Born, M. and Wolf, E. op cit p. 322

55a. Terman, F. E. op cit p. 39

55b. Fox, A.G. and Li, T. Bell System Tech. J. 40, 453 (1961)

56. Baumeister, P. W. and Jenkins, F. A. J. Opt. Soc. Am. 47, 57 (1957)

57. Terman, F. E. op cit p 42

58. Stone, J. M. Ph. D. Thesis, University of California, Berkeley, 1953 (unpublished)

59. Mielenz, K. D. J. Opt. Soc. Am. 50, 1014 (1960)

60. Smith, S. D. J. Opt. Soc. Am. 48, 43 (1958)

61. Baumeister, P. W., Jenkins, F. A. and Jeppesen, M. A. J. Opt. Soc. Am. 49, 1188 (1959)

62. Geffcken, W. agnew. Chem., A, 60 1 (1948)

63. Geffcken, W. Deutsches Reich Pat. #716,153 (Dec. 1939)

64. Turner, A. F. J. phys. radium 11, 444 (1950)

65. Korolev, F. A., Klement'eva, A. Yu., and Meshcheryakova, T. F. Optics and Spectroscopy (translation of Optika i Spectroskopia) 6, 341 (1960)

66. Turner, A. F. and Ullrich, O. A. J. Opt. Soc. Am. 38, 662 (A) (1948)

67. Mann, A. E. and Rock, F. C. J. Opt. Soc. Am. 38, 280 (A) (1958)

68. Berning, P. H. and Turner, A. F. J. Opt. Soc. Am. 47, 230 (1957)

69. Hadley, L. N. and Dennison, D. M. J. Opt. Soc. Am. 38, 483 (1948)

70. Ring, J., Beer, R., and Hewison, V., J. phys. radium 19, 321 (1958)

71. Dobrowolski, J. J. Opt. Soc. Am. 49, 794 (1959)

72a Lissberger, P. H. J. Opt. Soc. Am. 49, 121 (1959)

72b Lissberger, P. H. and Wilcock, W. L. J. Opt. Soc. Am. 49, 126 (1959)

72c Hass, G. and Salzburg, C. D. J. Opt. Soc. Am. 44, 181 (1954)

73. J. phys. radium 11, 305-480 (July 1950)

74. Vasicek, A. Optics of Thin Films (North-Holland, 1960)

75. Naval Ordnance Pamphlet (OP) 1952, Optics Filming, U.S. Gov. Printing Office, (1945)

21 COATING OF OPTICAL SURFACES

21.1 INTRODUCTION

21.1.1 Uses. Thin films of dielectric, metallic or even semi-conducting materials are most often applied to optical components such as lenses, plates and reflectors for the purpose of altering their energy reflectances or transmittances. A great variety of distributions of spectral reflectances and transmittances can be achieved over the ultraviolet, visible and infra-red regions. However, the number of materials having suitable optical, mechanical and chemical properties for use in the ultraviolet region is severely limited. Occasionally, thin films are used for modifying, especially at oblique incidence, phase changes as well as amplitude changes upon reflection. Thin films can also serve as protective coatings for surfaces of soft materials such as aluminum or silver. In another type of application, films are deposited with non-uniform thickness in order to achieve a slight degree of aspherization of the coated surface or in order to produce wedges that transmit non-uniformly in a specified manner. In still another broad class of films, a combination of optical properties such as transmittance, and of electrical properties such as conductance is provided. A diversity of specialized films consisting of combinations of two or more materials in a multilayer is required to meet an increasing list of modern applications.

21.1.2 Properties of thin films. We shall be concerned herein with the physical principles governing the optical properties of thin films. Fortunately, thin films have been found to behave to a good first approximation as homogeneous, plane-parallel layers that can be regarded as infinite in lateral dimensions. This idealized model of single films or multilayers can be analyzed without further approximation as a boundary problem involving Maxwell's equations. Several related forms of this useful theory will be treated. Actually, thin films are not homogeneous either laterally or along the thickness direction. Small departures from the predictions of the idealized model are therefore likely to occur. Theories dealing with inhomogeneity along the thickness direction are under active investigation; but it must be expected that these theories will be of greatest value in designing films whose inhomogeneities are increased deliberately.

21.2 DEFINITIONS AND PRINCIPLES

21.2.1 The optical constants. The optical constants, n and K, of an homogeneous, isotropic film are defined in the following manner. We take the solutions for the electric vector, E, and the magnetic vector, H, in the form

$$E = U e^{-i\omega t} \; ; \; H = V e^{-i\omega t} \; ; \qquad (1)$$

in which U and V are vectors; $U = (U_x, U_y, U_z)$ and $V = (V_x, V_y, V_z)$. Maxwell's curl relations become

$$\mathrm{Curl}\ V + i \frac{\omega}{c} m^2 U = 0; \qquad (2)$$

$$\mathrm{Curl}\ U - i \frac{\omega}{c} \mu V = 0; \qquad (3)$$

and the wave equations for U and V become

$$\nabla^2 U + \frac{\omega^2}{c^2} \mu m^2 U = 0; \qquad (4)$$

$$\nabla^2 V + \frac{\omega^2}{c^2} \mu m^2 V = 0; \qquad (5)$$

wherein

$$m^2 = \epsilon + i\, 4\pi\sigma/\omega \qquad \text{(defining m);} \qquad (6)$$

$$m = n(1 + iK) \qquad \text{(defining n and K).} \qquad (7)$$

The magnetic permeability, μ, and the dielectric constant, ϵ, are defined so that they are unity for vacuum. σ is the electric conductivity; c is velocity in vacuum and $\omega = 2\pi/T$, where T is the period of vibration of the wave. $i = \sqrt{-1}$.

$$\omega/c = 2\pi/\lambda_o = k \qquad (8)$$

where λ_o denotes wavelength in vacuum. As so defined, n and K are the optical constants that are usually listed in handbooks. In most optical problems one will take $\mu = 1$. We shall regard n and K as the fundamental properties that are measured from experimental observation on the wave motion. Under this view, ϵ and σ are derived from knowledge of n and K.

21. 2. 2 Physical significance. So much confusion exists about the physical significance of the optical constants, n and K, that further discussion is warranted. It is often believed that n is the refractive index such that the phase velocity v is always given by c/n. Suppose that a plane wave is propagated along Z with its electric vector polarized to vibrate along X. Then $U = (U_x, 0, 0)$. Since the wave is plane, U_x is, by assumption, independent of x and y. Hence the wave equation reduces to

$$\frac{\partial^2 U_x}{\partial z^2} + \frac{\omega^2}{c^2}\mu m^2 U_x = 0. \qquad (9)$$

It is easily verified by substitution into (9) that a solution is

$$U_x = E_o\, e^{i\frac{\omega}{c}\mu^{1/2} mz} = E_o\, e^{-k\mu^{1/2} nKz}\, e^{i\frac{\omega}{c}\mu^{1/2} nz} \qquad (10)$$

One concludes from equations (1) and (10) that

$$E = (1, 0, 0)\; E_o\, e^{-k\mu^{1/2} nKz}\; e^{-i\omega(t - \frac{z}{v})} \qquad (11)$$

wherein

$$v = c/\mu^{1/2} n \qquad (12)$$

is the phase velocity of the wave in the medium, and $k\mu^{1/2} n K$ is an absorption or extinction coefficient that determines the rate of attenuation of the amplitude with increasing z. We observe from equation (11) that the parallel planes, z = constant, are planes of equiphase (wavefronts) and of equiamplitude. When the planes of equiphase and equiamplitude are parallel, the wave is said to be homogeneous. If, then, $\mu = 1$ and the wave is homogeneous, the optical constant, n, is in fact the refractive index. When a homogeneous wave is incident normally upon any system of plane parallel layers, the wave remains homogeneous. But when a homogeneous wave is incident obliquely upon a system of plane parallel layers, the wave becomes inhomogeneous in the absorbing layers, i.e. planes of equiphase and equiamplitude do not remain strictly parallel when $K \neq 0$. With inhomogeneous waves, the phase velocity is not given by equation (12) and rate of attenuation is not governed by the product $k\mu^{1/2} n$ K. A generalized form of Snell's law of refraction applies, but the actual refractive index is not the optical constant n even when $\mu = 1$. It will not be an objective of this text to dwell upon the effects exhibited by inhomogeneous waves in absorbing media; but it is worth noting that the following systems of equations will include the effects produced by inhomogeneous waves in the absorbing layers. When the system is free of absorption, the waves remain homogeneous even at oblique incidence.

21. 2. 3 Fresnel's coefficients for normal incidence.

21. 2. 3. 1 Fresnel's coefficients of reflection and transmission with respect to a plane interface between two homogeneous media can be derived with a high degree of rigor. Derivations based upon Maxwell's equations and realistic boundary conditions will be found in almost all textbooks dealing with physical optics or electromagnetic theory. The following Fresnel coefficients apply to normal incidence upon the interface as illustrated in Figure 21. 1. The Z-direction is chosen normal to the interface, and the point z = 0 shall fall at the interface. To date, the permeabilities, μ, have invariably been unity in the optical applications of thin films. In view of the unlikelihood of cases $\mu \neq 1$, the following considerations will be restricted to cases $\mu = 1$.

21. 2. 3. 2 Let τ_o be a complex number that specifies the amplitude and phase of the incident E-vector at the left hand boundary, z = 0. Let r_o be a complex number that specifies the amplitude and phase of the reflected E-vector at the left hand boundary, z = 0. Similarly, let τ_1 specify the amplitude and phase of the transmitted E-vector at the right hand boundary, z = 0. It can be shown that

$$\frac{r_o}{\tau_o} = \frac{m_o - m_1}{m_o + m_1} \equiv W_1 , \qquad (13)$$

and that

$$\frac{\tau_1}{\tau_o} = \frac{2m_o}{m_o + m_1} , \qquad (14)$$

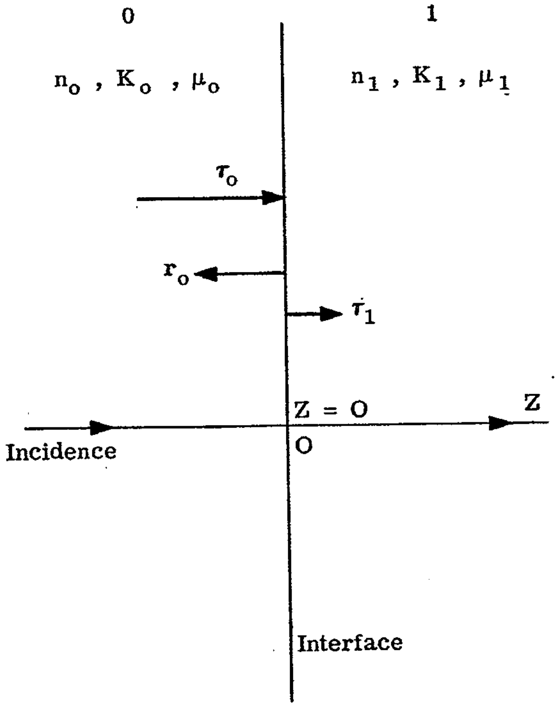


Figure 21. 1- Notation with respect to Fresnel's coefficients for normal incidence.

in which

$$m_j = n_j (1 + i K_j) ; \quad j = 0 \text{ and } 1. \tag{15}$$

The ratios r_o/τ_o and τ_1/τ_o are, respectively, Fresnel's coefficient of reflection and transmission at normal incidence. One may take τ_o as unity without any essential loss of generality.

21. 2. 3. 3 If neither medium is absorbing, $K_o = K_1 = 0$. Hence for interfaces between non-absorbing media the Fresnel coefficients reduce to the better known results

$$r_o/\tau_o = \frac{n_o - n_1}{n_o + n_1} , \tag{13a}$$

and

$$\tau_1/\tau_o = \frac{2n_o}{n_o + n_1} , \tag{14a}$$

in which n_o and n_1 are physically the refractive indices of the two media. When $n_1 > n_o$, one may write

$$r_o/\tau_o = \frac{|n_o - n_1|}{n_o + n_1} \quad e^{\pm i\pi} . \tag{13b}$$

Thus one concludes that, with respect to the electric vector, the phase change on reflection is ± π radians when $n_1 > n_o$. The phase change on transmission across the interface is always zero when the two media are non-absorbing, for then the ratio τ_1/τ_o is necessarily real and positive.

21. 2. 3. 4 As a second example, consider incidence from a non-absorbing medium upon an absorbing medium. Since $m_o = n_o$ and $m_1 = n_1 + i n_1 K_1$, one obtains from equation (13)

$$\frac{r_o}{\tau_o} = \frac{n_o - n_1 - i n_1 K_1}{n_o + n_1 + i n_1 K_1} = \frac{n_o^2 - n_1^2 (1 + K_1^2) - i\, 2 n_o n_1 K_1}{(n_o + n_1)^2 + n_1^2 K_1^2} . \tag{16}$$

For reflection from air to metals, $n_1^2 (1 + K_1^2)$ invariably exceeds n_o^2 . Hence the real and imaginary parts of the Fresnel reflection coefficient r_o/τ_o will usually be negative and the phase angle on reflection will fall in the third quadrant. Thus with $r_o/\tau_o = |r_o/\tau_o|\, e^{i\theta}$, $180^\circ < \theta < 270^\circ$. From equation (14),

$$\frac{\tau_1}{\tau_o} = \frac{2n_o}{n_o + n_1 + in_1 K_1} = 2\,\frac{n_o(n_o + n_1) - in_o n_1 K_1}{(n_o + n_1)^2 + n_1^2 K_1^2}\ . \qquad (17)$$

The phase angle introduced by transmission through the interface will fall in the fourth quadrant.

21. 2. 4 Fresnel's coefficients for oblique incidence.

21. 2. 4. 1 The direction, Z, has been taken along the normal to the interface. It is convenient to choose the X-direction in the plane of incidence as illustrated in Figure 21.2. X, Z is then the plane of incidence. Introduce for brevity,

$$p_o = \sin i_o\,; \quad q_o = \cos i_o\,; \qquad (18)$$

where i_o is the angle of incidence. Let

$$M_\nu = (m_\nu^2 - m_o^2 p_o^2)^{1/2} \qquad (19)$$

wherein the suffix $\nu = 0,\ 1$ and refers to the media of Figure 21. 2, and wherein m_ν is defined by equation (15). M_ν is complex imaginary whenever m_o or m_1 is complex imaginary. It can happen, as in total internal reflection, that $(m_\nu^2 - m_o^2 p_o^2)$ is real and less than zero. In such cases

$$M_\nu = i\,|m_\nu^2 - m_o^2 p_o^2|^{1/2}\,; \quad i = \sqrt{-1}\ . \qquad (19a)$$

21. 2. 4. 2 We need to introduce two more quantities W_ν and F_ν. These are

$$W_\nu = \frac{M_{\nu-1} - M_\nu}{M_{\nu-1} + M_\nu}\ , \qquad (20)$$

and

$$F_\nu = \frac{m_\nu^2 M_{\nu-1} - m_{\nu-1}^2 M_\nu}{m_\nu^2 M_{\nu-1} + m_{\nu-1}^2 M_\nu}\ . \qquad (21)$$

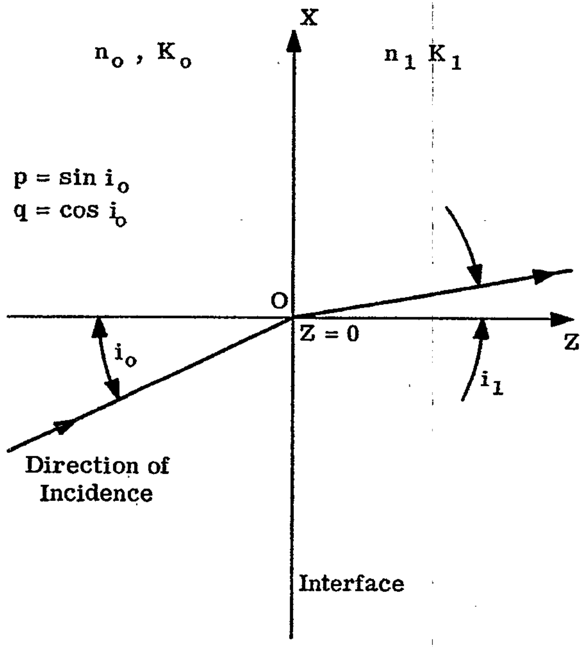


Figure 21. 2- Notation with respect to Fresnel's coefficients for oblique incidence. Plane XZ is chosen as the plane of incidence. i_o is the angle of incidence.

The laws of reflection depend upon the state of polarization of the incident wave. It suffices to consider the Fresnel coefficients of reflection and transmission for two states of polarization. In one of these states, the electric vector is perpendicular to the plane of incidence so that $E = (0, E_y, 0)$. In the second state of polarization, the magnetic vector is perpendicular to the plane of incidence so that $H = (0, H_y, 0)$. A minimum number of four Fresnel coefficients becomes necessary.

21.2.4.3 Consider first the state of polarization in which the electric vector is perpendicular to the plane of incidence. Let τ_o and r_o be complex numbers that specify the amplitude and phase of the incident and reflected electric vector at the origin, O, at left hand boundary of the interface, $z = 0$, Figure 21.2. Let, similarly, τ_1 specify the amplitude and phase of the transmitted electric vector at the point $x = z = 0$ at the right hand boundary of the interface. The ratios r_o/τ_o and τ_1/τ_o are now Fresnel's coefficients of reflection and transmission, respectively, for the electric vector. These ratios are given by

$$\frac{r_o}{\tau_o} = W_1 = \frac{M_o - M_1}{M_o + M_1} , \tag{22}$$

and

$$\frac{\tau_1}{\tau_o} = \frac{2M_o}{M_o + M_1} . \tag{23}$$

The similarity of equations (13) and (22), and of equations (14) and (23) should be noted. These results become alike when $p_o = 0$ (normal incidence).

21.2.4.4 Consider next the state of polarization in which the magnetic vector is perpendicular to the plane of incidence. Let T_o and R_o be complex numbers that specify the amplitude and phase of the incident and reflected H-vectors, respectively, at point $x = 0$ at the left hand boundary of the interface at $z = 0$. Let T_1 specify the amplitude and phase of the transmitted H-vector at point $x = 0$ at the right hand boundary of the interface. The ratios R_o/T_o and T_1/T_o define Fresnel's coefficient of reflection and transmission, respectively, for the perpendicular component of the magnetic vector. These ratios are given by

$$\frac{R_o}{T_o} = F_1 = \frac{m_1^2 M_o - m_o^2 M_1}{m_1^2 M_o + m_o^2 M_1} , \tag{24}$$

and

$$\frac{T_1}{T_o} = \frac{2m_1^2 M_o}{m_1^2 M_o + m_o^2 M_1} . \tag{25}$$

21.2.4.5 As an application or test of equations (22) and (24), consider total internal reflection. This phenomenon occurs when neither medium absorbs, so that $m_o = n_o$ and $m_1 = n_1$, and when $n_o > n_1$. Total internal reflection occurs when the angle of incidence $i_o \geq \sin^{-1} \frac{n_1}{n_o}$, i.e. when $n_o^2 p_o^2 \geq n_1^2$. Therefore, from equation (19a) it follows that M_1 is a pure imaginary number for angles i_o beyond the critical angle for total internal reflection. Since m_o, m_1 and M_o are real numbers, the numerators of equations (22) and (24) are complex conjugates with respect to the denominators. Hence it follows at once that both $|r_o/\tau_o|^2$ and $|R_o/T_o|^2$ are unity. The energy reflectance is therefore total, as required by experiment. One can verify that the numerator of equation (24) is zero at Brewster's angle. This means that the H-vector will not be reflected when it is perpendicular to the plane of incidence, or, equivalently, that the E-vector will not be reflected at Brewster's angle when it vibrates in the plane of incidence.

21.2.4.6 Whereas $|r_o/\tau_o|^2$ and $|R_o/T_o|^2$ can always be interpreted as energy reflectances for the states of polarization to which they apply, the square of the absolute values of Fresnel's coefficients of transmission τ_1/τ_o and T_1/T_o do not necessarily equal the actual energy transmittances. This matter will be treated in some detail since it has been the source of much confusion. With respect to thin films or interfaces, one invariably wishes to compute energy transmittance for cases in which the initial and last media are non-absorbing. Accordingly, emphasis will be placed upon non-absorbing initial and final media.

21.2.5 The electromagnetic field when the electric vector is perpendicular to the plane of incidence.

21.2.5.1 Let us suppose that an homogeneous incident wave has been generated in the initial medium such that the incident electric vector is the known vector,

$$E_{\text{incident}} = (0, 1, 0)\, \tau_o\, e^{-i\omega t}\, e^{ikm_o(p_o x + q_o z)} \tag{26}$$

From equations (1) and (26),

$$U_{\text{incident}} = (0,\ 1,\ 0)\ \tau_o\ e^{ikm_o(p_o x + q_o z)} \qquad (26a)$$

Curl relation, equation (3), serves to determine V from U. For plane waves incident in the X, Z plane, vectors U and V are independent of y. Consequently equation (3) yields directly the result,

$$V_x = \frac{i}{k\mu}\ \frac{\partial U_y}{\partial z}\ , \quad V_y = 0, \quad V_z = \frac{-i}{k\mu}\ \frac{\partial U_y}{\partial x}\ ; \qquad (27)$$

a result that holds in any medium. Since U_y is given from equation (26a), one can compute vector, V, from equation (27) and then write the incident H-vector from equation (1). One obtains in a straightforward manner for the case $\mu = 1$,

$$H_{\text{incident}} = (-q_o,\ 0,\ p_o)\ m_o\ \tau_o\ e^{-i\omega t}\ e^{ikm_o(p_o x + q_o z)}\ , \qquad (26b)$$

The incident electromagnetic field becomes known when τ_o is assigned. The time averaged Poynting vector, S, is given except for an unimportant factor by the vector product

$$S = E \times \bar{H} + \bar{E} \times H\ . \qquad (28)$$

From equations (26), (26b) and (28),

$$S_{\text{incident}} = (p_o,\ 0,\ q_o)\ 2n_o\,|\tau_o|^2\ e^{-2kn_o K_o(p_o x + q_o z)}\ , \qquad (26c)$$

since $m_o + \bar{m}_o = 2n_o$. This energy flux is along the direction of propagation of the incident wave.

21.2.5.2 The incident E-vector is reflected with the Fresnel reflection coefficient r_o/τ_o. Consequently,

$$E_{\text{reflected}} = (0,\ 1,\ 0)\ \frac{r_o}{\tau_o}\ \tau_o\ e^{-i\omega t}\ e^{ikm_o(p_o x - q_o z)}\ . \qquad (29)$$

Just as equation (27) served for determining H_{incident} from E_{incident}, it serves again for determining $H_{\text{reflected}}$ from $E_{\text{reflected}}$. One obtains straightforwardly,

$$H_{\text{reflected}} = (q_o,\ 0,\ p_o)\ m_o\ \frac{r_o}{\tau_o}\ \tau_o\ e^{-i\omega t}\ e^{ikm_o(p_o x - q_o z)}\ . \qquad (29a)$$

The Fresnel coefficient r_o/τ_o of equation (22) determines the reflected electromagnetic field. Upon evaluating the Poynting vector, S, from equations (28), (29) and (29a), one obtains

$$S_{\text{reflected}} = (p_o,\ 0,\ -q_o)\ 2n_o \left|\frac{r_o}{\tau_o}\right|^2 |\tau_o|^2\ e^{-2kn_o K_o(p_o x - q_o z)}\ , \qquad (29b)$$

an energy flux along the direction of the reflected wave.

21.2.5.3 The electric vector transmitted across the interface $z = 0$, Figure 21.2, has the form*

$$E_{\text{transmitted}} = (0,\ 1,\ 0)\left(\frac{\tau_1}{\tau_o}\right)\tau_o\ e^{-i\omega t}\ e^{ik(m_o p_o x + M_1 z)}\ . \qquad (30)$$

Correspondingly, equation (27) yields

$$H_{\text{transmitted}} = (-M_1,\ 0,\ m_o p_o)\left(\frac{\tau_1}{\tau_o}\right)\tau_o\ e^{-i\omega t}\ e^{ik(m_o p_o x + M_1 z)}\ . \qquad (30a)$$

The wave described by equations (30) and (30a) is in general inhomogeneous.** It is more convenient for many

---

*It can be verified easily by substitution, that vector U defined by equations (1) and (30) satisfies wave equation (4) in medium number one, provided that M_1 obeys equation (19).

**Suppose, for example, that $m_o = n_o$ but that m_1 is complex. The first medium is then non-absorbing and the second medium is absorbing. M_1 is now complex imaginary. Write $M_1 = R_e(M_1) + i\,I_m(M_1)$. Since

$$e^{ik(m_o p_o x + M_1 z)} = e^{-kI_m(M_1)z}\ e^{ik[n_o p_o x + R_e(M_1) z]}$$

the planes of equiamplitude are parallel to the interface $z = 0$ whereas the planes of equiphase are the planes $n_o p_o x + R_e(M_1)\,z$ = constant.

purposes to write the second exponential in equations (30) and (30a) in the expanded form,

$$e^{ik(m_o p_o x + M_1 z)} = e^{-k[n_o K_o p_o x + I_m(M_1) z]} \; e^{ik[n_o p_o x + R_e(M_1) z]} , \qquad (30b)$$

wherein $R_e(M_1)$ and $I_m(M_1)$ denote, respectively, the real and imaginary parts of M_1. The Poynting vector, S, can now be found in a straightforward manner by applying equation (28) to equations (30), (30a) and (30b). The most general result is

$$S_{\text{transmitted}} = \left[n_o p_o,\ 0,\ R_e(M_1)\right] \; 2 \left|\frac{\tau_1}{\tau_o}\right|^2 |\tau_o|^2 \; e^{-2k[n_o K_o p_o x + I_m(M_1) z]} . \qquad (30c)$$

By comparing the three components $n_o p_o$, 0 and $R_e(M_1)$ with the arguments $n_o p_o x + R_e(M_1) z$ of the second exponential of equation (30b), one finds that $S_{\text{transmitted}}$ is an energy flow along the direction of propagation of the equiphase surfaces (wavefronts) in the second medium when the electric vector is perpendicular to the plane of incidence. Equation (30c) becomes most significant and useful when the initial medium has negligible absorption so that $K_o = 0$; for then the Poynting vector is constant in planes z = constant. Explicitly,

$$S_{\text{transmitted}} = \left[n_o p_o,\ 0,\ R_e(M_1)\right] \; 2 \left|\frac{\tau_1}{\tau_o}\right|^2 |\tau_o|^2 \; e^{-2k I_m(M_1) z} . \qquad (30d)$$

If, also, the second medium is non-absorbing, $m_1 = n_1$ and M_1 is real. Attenuation with z does not occur because $I_m(M_1)$ is zero.

21.2.5.4 Since the x - and z - components of $S_{\text{transmitted}}$ are proportional to $n_o p_o$ and $R_e(M_1)$, respectively, the angle, i_1, between $S_{\text{transmitted}}$ and the Z-axis is given by

$$\tan i_1 = n_o p_o / R_e(M_1) , \qquad (30e)$$

or

$$\sin i_1 = n_o p_o \Big/ \left[n_o^2 p_o^2 + R_e^2(M_1)\right]^{1/2} . \qquad (30f)$$

As stated, this direction of the Poynting vector, S, is parallel to the direction of propagation of the wavefronts. Let

$$(n_a)_1 = \left[n_o^2 p_o^2 + R_e^2(M_1)\right]^{1/2} . \qquad (31)$$

Equation (30f) now states that the law of refraction is

$$(n_a)_1 \sin i_1 = n_o p_o = n_o \sin i_o \qquad (32)$$

in which $(n_a)_1$ is in fact the actual refractive index of medium number one, Figure 21.2. When $m_o = n_o$ and $m_1 = n_1$, $(n_a)_1 = n_1$ so that the more general law of equation (32) degenerates into the usual form known as Snell's law. Equation (32) serves to determine the direction of the Poynting vector and the "rays" in medium number one.

21.2.5.5 The manner in which the actual refractive index, $(n_a)_1$, depends upon the angle of incidence, i_o, and upon $n_1 K_1$, is described by the table in Table 21.1 for the case in which $n_o = 1$, $K_o = 0$, and $n_1 = 1.75$. $(n_a)_1$ increases with the angle of incidence and with the product $n_1 K_1$. Because nK is less than 0.02 in the usual lenses, plates, etc., the more general law of equation (32) is not of great importance to geometrical optics.

21.2.5.6 The absolute value of the vector $n_o p_o$, 0, $R_e(M_1)$ in equation (30d) is $n_o^2 p_o^2 + R_e^2(M_1)$ $^{1/2}$. Hence equation (30d) can be written in the more useful and significant form

$$\left|S_{\text{transmitted}}\right| = 2\,(n_a)_1 \left|\frac{\tau_1}{\tau_o}\right|^2 |\tau_o|^2 \; e^{-2k I_m(M_1) z} . \qquad (33)$$

The relations between the Fresnel coefficients, r_o/τ_o and τ_1/τ_o, and the energy reflectance and energy transmittance, respectively, can now be clarified unambiguously. First, we note from equations (26c) and (29b) that at $z = 0$,

$$\frac{\left|S_{\text{reflected}}\right|}{\left|S_{\text{incident}}\right|} = \left|\frac{r_o}{\tau_o}\right|^2 \quad \text{(energy reflectance)} , \qquad (34)$$

a result that holds whether or not m_o is complex. Secondly, we note from equations (33) and (26c), that

$n_1 K_1$ \ i_o	0°	10°	20°	40°	60°	80°
0.01	1.750000	1.750000	1.750001	1.7500045	1.750009	1.750013
0.02	1.750000	1.750001	1.7500045	1.750018	1.750037	1.750053
0.04	1.750000	1.7500045	1.750015	1.750071	1.750148	1.750212
0.06	1.750000	1.750010	1.750041	1.750160	1.750333	1.750493
0.08	1.750000	1.7500205	1.750072	1.750284	1.750591	1.750843
0.10	1.750000	1.750028	1.750113	1.750444	1.750921	1.751314
0.50	1.750000	1.750656	1.752605	1.760028	1.770224	1.778142
1.00	1.750000	1.752131	1.758389	1.781126	1.809765	1.830155
2.00	1.750000	1.754882	1.768962	1.817250	1.872569	1.908867
4.00	1.750000	1.757218	1.7778585	1.846750	1.922468	1.970524

Table 21.1- Table of values of the actual refractive index $(n_a)_1$ of the second medium as a function of the angle of incidence, i_o, and of $n_1 K_1$ for the case in which the first medium does not absorb and has the optical constant $n_o = 1$. The optical constant $n_1 = 1.750000$.

at $z = 0$,

$$\frac{|S_{\text{transmitted}}|}{|S_{\text{incident}}|} = \frac{(n_a)_1}{n_o}\left|\frac{\tau_1}{\tau_o}\right|^2 \tag{35}$$

a result that holds when $m_o = n_o$. Consider next the conservation of energy flow across any element of area, ΔA, of the interface $z = 0$, Figure 21.3. Energy is conserved provided that

$$|S_{\text{reflected}}|\ \Delta A_r + |S_{\text{transmitted}}|\ \Delta A_t = |S_{\text{incident}}|\ \Delta A_i\ , \tag{36}$$

in which the elements of area are interrelated as indicated in Figure 21.3. Division of equation (36) by the right hand member produces the following important result,

$$\frac{|S_{\text{reflected}}|}{|S_{\text{incident}}|}\frac{\Delta A_r}{\Delta A_i} + \frac{|S_{\text{transmitted}}|}{|S_{\text{incident}}|}\frac{\Delta A_t}{\Delta A_i} = 1. \tag{37}$$

The first left hand member is energy reflectance from the element of area ΔA. Because $\Delta A_r = \Delta A_i$, comparison of the first left hand member of equations (37) and (34), shows that the ratio $|r_o/\tau_o|^2$ is in fact energy reflectance. The second left hand member of equation (37) is energy transmittance of the element of area ΔA. Since $\Delta A_t/\Delta A_i = \cos i_1/\cos i_o$, equations (37) and (35) show that the energy transmittance of any element ΔA of the interface, $z = 0$, is given by

$$\text{Energy transmittance} = \frac{(n_a)_1}{n_o}\left|\frac{\tau_1}{\tau_o}\right|^2 \frac{\cos i_1}{\cos i_o}\ , \tag{38}$$

in which $(n_a)_1$ is the actual refractive index of the last medium, and τ_1/τ_o is Fresnel's transmission coefficient for the case in which the E-vector is perpendicular to the plane of incidence.

21.2.5.7 We shall verify that equations (22) and (23) for the Fresnel coefficients are consistent with the law of conservation of energy. Upon introducing equations (34) and (35) into equation (37), one obtains the condition for conservation of energy at the interface $z = 0$ in the form

$$\left|\frac{r_o}{\tau_o}\right|^2 + \frac{(n_a)_1 \cos i_1}{n_o \cos i_o}\left|\frac{\tau_1}{\tau_o}\right|^2 = 1. \tag{39}$$

Hence one should obtain

$$\frac{|M_o - M_1|^2}{|M_o + M_1|^2} + \frac{(n_a)_1 \cos i_1}{n_o \cos i_o}\ \frac{4|M_o|^2}{|M_o + M_1|^2} = 1. \tag{40}$$

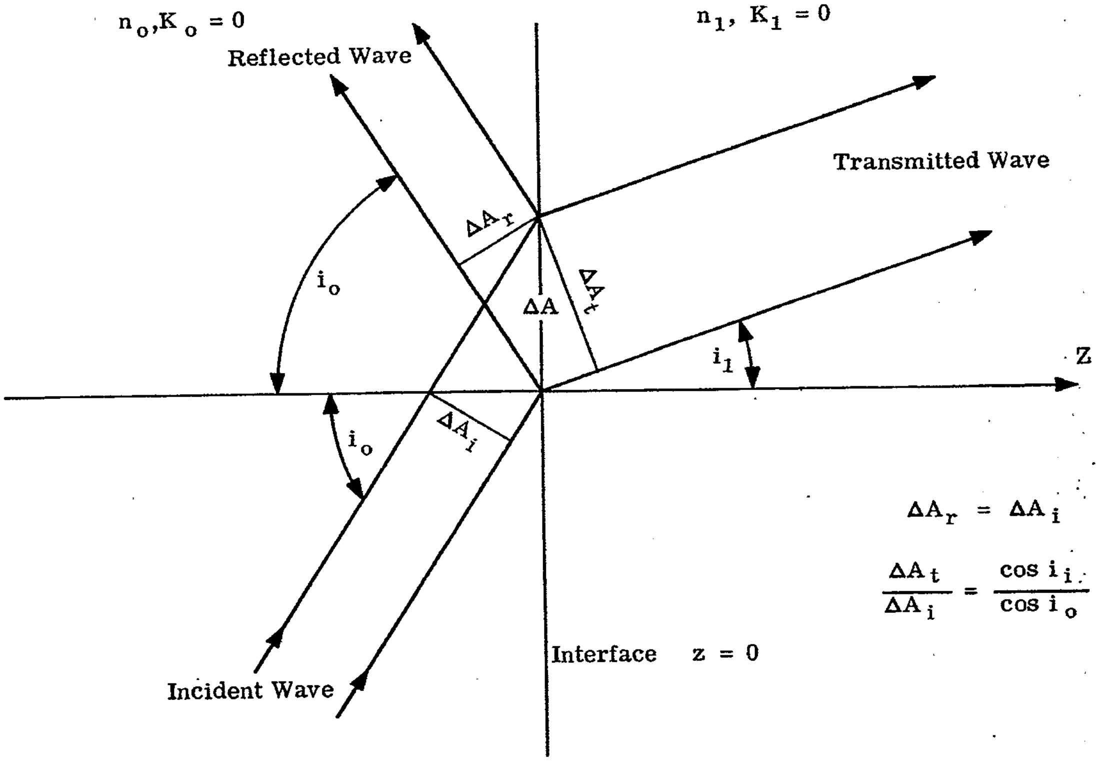


Figure 21. 3- Notation with respect to the flow of energy flux in the Poynting vectors for cases in which absorption is negligible in the initial and final media.

To avoid difficulties, let us test the case $m_o = n_o$ and $m_1 = n_1$. In this case $n_o \cos i_o = M_o$ and $(n_a)_1 \cos i_1 = M_1$. Hence equation (40) becomes the required identity.

21. 2. 6 The electromagnetic field when the magnetic vector is perpendicular to the plane of incidence.

21. 2. 6. 1 When the y-component of the H-vector is given, the curl relation, equation (2), determines the vector U. In the present case

$$U_x = \frac{-i}{km^2} \frac{\partial V_y}{\partial z} \quad ; \quad U_y = 0 \; ; \quad U_z = \frac{i}{km^2} \frac{\partial V_y}{\partial x} \, . \tag{41}$$

Consistent with equation (26), we suppose that the incident magnetic vector is given as the homogeneous wave

$$H_{\text{incident}} = (0, 1, 0) \; T_o \; e^{ikm_o (p_o x + q_o z)} \, . \tag{42}$$

Then from equations (41) and (42),

$$E_{\text{incident}} = (q_o, 0, -p_o) \frac{T_o}{m_o} \; e^{ikm_o (p_o x + q_o z)} \, . \tag{42a}$$

Therefore,

$$S_{\text{incident}} = (p_o, 0, q_o) \frac{2 n_o}{|m_o|^2} \left| T_o \right|^2 e^{-2kn_o K_o (p_o x + q_o z)} \, . \tag{42b}$$

Since the Fresnel reflection coefficient is now R_o/T_o, equation (24),

$$H_{\text{reflected}} = (0, 1, 0) \left(\frac{R_o}{T_o} \right) T_o \; e^{ikm_o (p_o x - q_o z)} \, , \tag{43}$$

$$E_{\text{reflected}} = (-q_o, 0, -p_o) \frac{T_o}{m_o} \left(\frac{R_o}{T_o} \right) e^{ikm_o (p_o x - q_o z)} \, , \tag{43a}$$

and

$$S_{\text{reflected}} = (p_o, 0, -q_o) \frac{2 n_o}{|m_o|^2} (T_o)^2 \left| \frac{R_o}{T_o} \right|^2 e^{-2kn_o K_o (p_o x + q_o z)} \, . \tag{43b}$$

The incident and reflected Poynting vectors point along the direction of propagation of the corresponding waves.

21. 2. 6. 2 As in equation (30),

$$H_{\text{transmitted}} = (0,\ 1,\ 0) \left(\frac{T_1}{T_o}\right) T_o\ e^{ik(m_o p_o x + M_1 z)}, \tag{44}$$

therefore ,

$$E_{\text{transmitted}} = (M_1,\ 0,\ -m_o p_o) \frac{1}{m_1^2} \left(\frac{T_1}{T_o}\right) T_o\ e^{ik(m_o p_o x + M_1 z)}. \tag{44a}$$

When $m_o = n_o$,

$$S_{\text{transmitted}} = \left(n_o p_o\left[\frac{1}{m_1^2} + \frac{1}{\overline{m_1^2}}\right],\ 0,\ \frac{M_1}{m_1^2} + \frac{\overline{M}_1}{\overline{m_1^2}}\right) \left|\frac{T_1}{T_o}\right|^2 \left|T_o\right|^2 e^{-2k I_m (M_1) z}. \tag{44b}$$

Since

$$\frac{1}{m_1^2} + \frac{1}{\overline{m_1^2}} = \frac{m_1^2 + \overline{m_1^2}}{|m_1|^4} = \frac{2 R_e (m_1^2)}{|m_1|^4}, \tag{44c}$$

the x-component of S would change sign with $R_e (m_1^2)$. The possibility $R_e (m_1^2) < 0$ would violate also equation (6) because a negative dielectric constant, ϵ_1, has no physically acceptable meaning. Thus, measured values of n and K can be valid (1a) only when

$$R_e (m^2) = n^2 (1 - K^2) = \epsilon > 0. \tag{44d}$$

With respect to the listed values of n and K for metals, one generally finds that $K^2 > 1$ so that $n^2 (1 - K^2) < 0$. Application of the theory to cases in which the listed K-values exceed unity is to be regarded, therefore, with skepticism.(1) When $R_e (m_1^2) > 0$, one finds quite directly from equation (44b) that

$$S_{\text{transmitted}} = \left(n_o p_o,\ 0,\ R_e (M_1)\right) 2\ \frac{R_e (m_1^2)}{|m_1|^4} \left|\frac{T_1}{T_o}\right|^2 \left|T_o\right|^2 e^{-2k I_m (M_1) z}$$
$$+ (0,\ 0,\ 1)\ 2\ I_m (M_1) \frac{I_m (m_1^2)}{|m_1|^4} \left|\frac{T_1}{T_o}\right| \left|T_o\right|^2 e^{-2k I_m (M_1) z}. \tag{44e}$$

The first right hand vector is parallel to the normals to the equiphase surfaces (wavefronts) in the second medium. The second right hand vector is normal to the interface and vanishes as the imaginary part of m approaches zero. As $n_1 K_1$ is increased from zero, the transmitted Poynting vector, S, departs from the direction of the wave normals and tends toward perpendicularity with the interface when the incident H-vector is perpendicular to the plane of incidence.

21. 2. 6. 3 We shall restrict our considerations of equation (44e) to cases in which K_1 is so small that the vector with components (0, 0, 1) is negligible. The transmitted Poynting vector, S, and the wave normals are then practically parallel. As in 21. 2. 5, we obtain

$$|S_{\text{transmitted}}| = \frac{2 (n_a)_1}{|m_1|^4} R_e (m_1^2) \left|\frac{T_1}{T_o}\right|^2 \left|T_o\right|^2 e^{-2k I_m (M_1) z}, \tag{44f}$$

in which $(n_a)_1$ is given by equation (31) and in which S points along the direction determined by equation (32). From equations (42b) and (43b), one finds that at $z = 0$

$$\frac{|S_{\text{reflected}}|}{|S_{\text{incident}}|} = \left|\frac{R_o}{T_o}\right|^2 \quad \text{(energy reflectance)}, \tag{45}$$

a result that holds whether or not m_o is complex. One finds from equations (44f) and (42b) that at the interface, $z = 0$, with the approximation $R_e (m_1^2) = n_1^2$

$$\frac{|S_{\text{transmitted}}|}{|S_{\text{incident}}|} = \frac{(n_a)_1\ n_1^2}{|m_1|^4\ n_o} \left|\frac{T_1}{T_o}\right|^2, \tag{45a}$$

a result that has been specialized to the case $m_o = n_o$. Under the restriction that the transmitted Poynting vector is practically parallel to the wave normals in the second medium, equation (37) holds again. Hence

---

(1) See P. Drude pg. 368 Theory of Optics, Longmans Green & Co. 1902.

(1a). For a more modern viewpoint relative to cases in which $n^2(1-K^2) = \epsilon < 0$, consult the physical interpretation of negative dielectric constants by Max Born and Emil Wolf, Principles of Optics, Pergamon Press, (1959), pp 618 and 623.

$R_o/T_o\ ^2$ is in fact energy reflectance. The steps leading to equation (38) now yield, instead of equation (38), the result,

$$\text{Energy transmittance} = \frac{(n_a)_1 \; n_1^2 \; n_o}{|m_1|^4} \left| \frac{T_1}{T_o} \right|^2 \frac{\cos i_1}{\cos i_o} , \tag{45b}$$

wherein T_1/T_o is Fresnel's coefficient of transmission of the H-vector when this vector is perpendicular to the plane of incidence. If $m_1 = n_1$, $(n_a)_1 = n_1$ and

$$\text{Energy transmittance} = \frac{n_o}{n_1} \left| \frac{T_1}{T_o} \right|^2 \frac{\cos i_1}{\cos i_o} , \tag{45c}$$

equations (45b) and (45c) should be compared with equation (38) for the state of polarization in which the E-vector is perpendicular to the plane of incidence. It will be seen that the ratios of the refractive indices are the inverse of one another. The Fresnel coefficients R_o/T_o and T_1/T_o of equations (24) and (25) are consistent with the requirement that the corresponding flow of energy shall be conserved at the interface $z = 0$.

21. 2. 7 Summary with respect to the Fresnel coefficients.

21. 2. 7. 1 The Fresnel coefficients r_o/τ_o and τ_1/τ_o given by equations (22) and (23), respectively, determine the reflected and transmitted E-vector when this vector is perpendicular to the plane of incidence. The corresponding incident, reflected and transmitted electromagnetic fields are given by equations (26), (29) and (30), respectively. Whereas $|r_o/\tau_o|^2$ is energy reflectance, the quantity $|\tau_1/\tau_o|^2$ is not, in general, energy transmittance. A study of the incident, reflected and transmitted Poynting vectors shows that energy transmittance across the interface between the two media is given by equation (38).

21. 2. 7. 2 The Fresnel coefficients R_o/T_o and T_1/T_o refer to reflection and transmission of the H-vector when it is perpendicular to the plane of incidence. With this state of polarization the incident, reflected and transmitted electromagnetic fields are determined by equations (42), (43) and (44). Examination of the time-averaged Poynting vectors shows that $|R_o/\tau_o|^2$ is energy reflectance but that $|T_1/T_o|^2$ is not necessarily energy transmittance. Equation (45b) is an approximate one for computing energy transmittance. As the absorption of the second medium vanishes, equation (45b) becomes more exact and approaches the result given by equation (45c) for the case in which neither medium is absorbing.

21. 2. 7. 3 When the first medium is non-absorbing and the second medium is absorbing, equation (31) and (32) show how the wave normals are refracted. This law of refraction is the same for both states of polarization. Whereas the time-averaged Poynting vector points along the wave normals when the E-vector is perpendicular to the plane of incidence, this Poynting vector does not always do so when the H-vector is perpendicular to the plane of incidence.

21. 2. 8 Normal incidence upon multilayers.

21. 2. 8. 1 Once the Fresnel coefficients for an interface between two media have been established, it is not necessary to resubmit the numerous interfaces of a multilayer to the Maxwell equations and the boundary conditions in order to construct the theory of thin films. Instead, the following instructive method can be utilized. We consider the useful case of normal incidence in order to simplify the presentation. We may suppose, without any essential loss of generality, that the electric vector is perpendicular to the plane X, Z, Figure 21. 4. In other words, we choose Z along the normal to the interfaces and take Y as the direction of vibration of the electric vector. Because the incident waves are assumed to be plane, the magnetic vector now vibrates along the X-direction.

21. 2. 8. 2 Waves propagated to the right and left in Figure 21. 4 are regarded as transmitted and reflected waves, respectively. In dealing with more than one interface, it is convenient to take τ_ν as a complex number that specifies the amplitude and phase of the "transmitted" electric vector at the right hand boundary of the ν^{th} medium or layer for the range of ν from 0 to N. As indicated in Figure 21. 4, τ_{N+1} specifies the amplitude and phase of the wave in the last medium $N + 1$ at the left hand boundary of this medium. Similarly, r_ν is a complex number that specifies the amplitude and phase of the reflected electric vector at the right hand boundary of the ν^{th} medium or layer. A reflected wave does not exist in the last medium because it extends indefinitely along Z. The ratio r_ν/τ_ν is complex reflectance at the right hand boundary of the ν^{th} medium or layer. We define

$$\rho_\nu \equiv r_\nu/\tau_\nu \quad \text{(complex reflectance).} \tag{46}$$

21. 2. 8. 3 The Fresnel coefficient of reflectance from medium number 0 to medium number 1 is given by $W_1 = (M_o - M_1)/(M_o + M_1)$ as in equation (22). More generally, the Fresnel coefficient of reflectance at the interface between the $\nu - 1^{th}$ and ν^{th} layer is given by W_ν, equation (20), when the light is incident

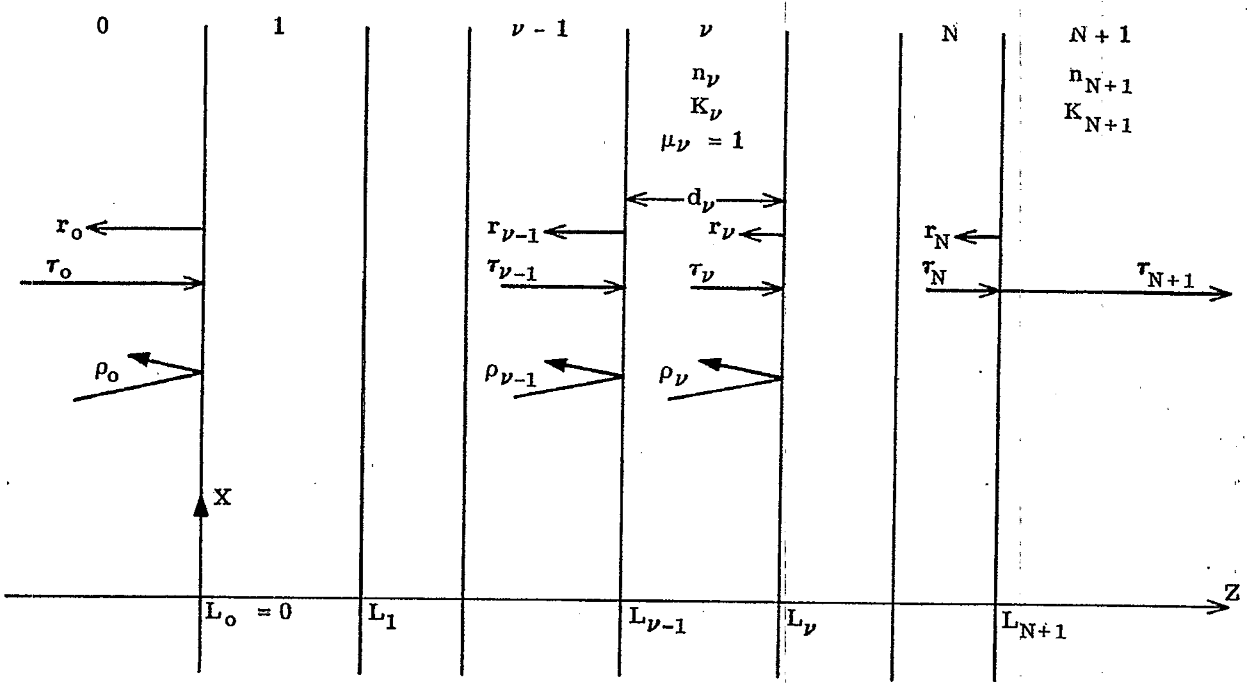


Figure 21. 4- Notation with respect to a system of N layers at normal incidence. The electric vector is taken perpendicular to the XZ plane. Incidence is from the O th medium upon layer #1. Both the initial and final medium #N + 1 are assumed to extend indefinitely along Z.

from the ν-1 th layer upon the ν th layer. When the direction of incidence is reversed, the effect is to interchange the integers ν and ν-1 in equation (20) so that Fresnel's coefficient of reflection becomes $-W_\nu$. In dealing with a system of layers it is therefore convenient to call W_ν the Fresnel coefficient of reflectance, when the electric vector is polarized to vibrate along the Y-direction. The ratios $\rho_\nu = r_\nu/\tau_\nu$ are equal to the Fresnel coefficients of reflectance only in special cases such as, for example, the single interface of Figure 21. 1.

21. 2. 8. 4 We have seen, as in equation (23), that Fresnel's coefficient of transmission across the interface from medium number 0 into medium number 1 is equal to $2M_0/(M_0 + M_1)$. More generally, Fresnel's coefficient of transmission across the interface from the $(\nu - 1)$ th into the ν th medium is equal to $2M_{\nu-1}/(M_{\nu-1} + M_\nu)$. Fresnel's coefficient of transmission through this interface in the opposite direction is equal to $2M_\nu/(M_{\nu-1} + M_\nu)$. For normal incidence $M_\nu = m_\nu = n_\nu(1 + i K_\nu)$ since $p_o = 0$ in equation (19). Let

$$\beta_\nu \equiv \frac{4\pi}{\lambda} m_\nu d_\nu = \frac{4\pi}{\lambda} n_\nu d_\nu + i \frac{4\pi}{\lambda} n_\nu K_\nu d_\nu , \qquad (47)$$

where d_ν is the thickness of the ν th layer. When $K_\nu = 0$, β_ν is twice the optical path (in radians) of the ν th layer.

21. 2. 8. 5 The following equilibrium theory for the flow of the electric vector through the multilayer can now be derived in a simple manner. We fix our attention upon the equilibrium flow at the $(\nu$-1) th and ν th layers as illustrated in Figure 21. 5. Consider the flow described by $r_{\nu-1}$. First, the flow $\tau_{\nu-1}$ is reflected at the interface at the right hand side of the $(\nu$-1) th layer in the direction of $r_{\nu-1}$ as the flow $\tau_{\nu-1} W_\nu$. Secondly, the flow r_ν arrives at the left side of the ν th layer as the flow $r_\nu e^{i\frac{\beta_\nu}{2}}$ and then passes through the interface subject to the Fresnel coefficient of transmission $2M_\nu/(M_{\nu-1} + M_\nu)$. Hence

$$r_{\nu-1} = \tau_{\nu-1} W_\nu + r_\nu e^{i\frac{\beta_\nu}{2}} \frac{2 M_\nu}{M_{\nu-1} + M_\nu} . \qquad (48)$$

Similarly,

$$\tau_\nu = \tau_{\nu-1} \frac{2 M_{\nu-1}}{M_{\nu-1} + M_\nu} e^{i\frac{\beta_\nu}{2}} - r_\nu W_\nu e^{i\beta_\nu} . \qquad (48a)$$

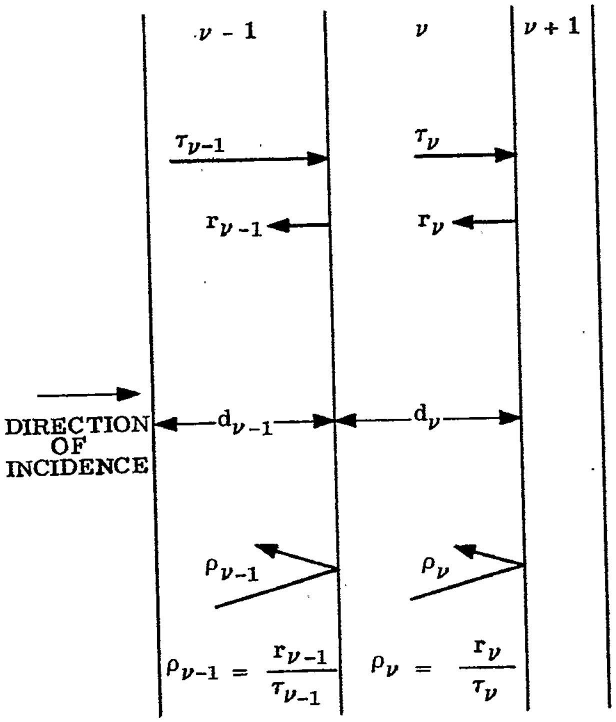


Figure 21. 5- Flow conditions at the $(\nu-1)^{th}$ and ν^{th} layers.

The second right hand member has the factor $e^{i\beta\nu}$ because the flow r_ν must cross the ν^{th} layer twice. Furthermore, the flow r_ν is reflected at the left hand interface of the ν^{th} layer so that the corresponding Fresnel coefficient of reflection is $-W_\nu$. From equation (48),

$$\frac{r_{\nu-1}}{\tau_{\nu-1}} = \rho_{\nu-1} = W_\nu + \frac{r_\nu}{\tau_{\nu-1}} e^{i\frac{\beta\nu}{2}} \frac{2 M_\nu}{M_{\nu-1} + M_\nu} . \tag{48b}$$

From equation (48a),

$$\tau_{\nu-1} = (\tau_\nu + r_\nu W_\nu e^{i\beta\nu}) \frac{M_{\nu-1} + M_\nu}{2 M_{\nu-1}} e^{-i\frac{\beta\nu}{2}} . \tag{48c}$$

By eliminating $\tau_{\nu-1}$ from equation (48b) with the aid of equation (48c), one obtains

$$\rho_{\nu-1} = W_\nu + \frac{r_\nu e^{i\beta\nu}}{\tau_\nu + r_\nu W_\nu e^{i\beta\nu}} \frac{4 M_{\nu-1} M_\nu}{(M_{\nu-1} + M_\nu)^2} . \tag{48d}$$

Dividing numerator and denominator in the right hand member by τ_ν yields the result

$$\rho_{\nu-1} = W_\nu + \frac{\rho_\nu e^{i\beta\nu}}{1 + W_\nu \rho_\nu e^{i\beta\nu}} \frac{4 M_{\nu-1} M_\nu}{(M_{\nu-1} + M_\nu)^2} . \tag{48e}$$

From equation (48e),

$$\rho_{\nu-1} = \frac{\rho_\nu e^{i\beta\nu} \left[W_\nu^2 + 4M_{\nu-1} M_\nu/(M_{\nu-1} + M_\nu)^2\right] + W_\nu}{1 + W_\nu \rho_\nu e^{i\beta\nu}} . \tag{48f}$$

It follows directly from the definition of W_ν, equation (20), that

$$W_\nu^2 + 4M_{\nu-1} M_\nu/(M_{\nu-1} + M_\nu)^2 = 1 . \tag{49}$$

Hence

$$\rho_{\nu-1} = \frac{\rho_\nu \, e^{i\beta\nu} + W_\nu}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \, , \qquad (50)$$

a recursion formula that enables one to compute $\rho_{\nu-1}$ from ρ_ν and the given properties β_ν and W_ν of the νth layer. Equation (50) is the well known result that follows rigorously from the Maxwell equations and the boundary conditions appropriate to a multilayer.

21.2.8.6 The method for computing the complex reflectance ρ_o of the entire multilayer is now clear. Since $\rho_{N+1} = 0$ because no reflected wave exists in the final medium, number N + 1, it follows from equation (50) that

$$\rho_N = W_{N+1} = \frac{M_N - M_{N+1}}{M_N + M_{N+1}} \, . \qquad (51)$$

This result is to be expected because ρ_N ought to be the Fresnel coefficient of reflection at the last interface of the multilayer. Since ρ_N becomes known, one computes ρ_{N-1} from equation (50). This equation is then applied consecutively to determine $\rho_{N-2}, \rho_{N-3} \cdots\cdots\cdots \rho_o$. If ρ_o is expressed in the form

$$\rho_o = |\rho_o| \; e^{i\theta_o} \, , \qquad (52)$$

then $|\rho_o|$ is amplitude reflectance and θ_o is phase change on reflection of the entire multilayer at the first interface $z = 0$, Figure 21.4. The phase change on reflection is phase retardation (equivalent to an increase in optical path when $\theta_o > 0$).

21.2.8.7 The complex transmittance τ_{N+1}/τ_o of the system of layers is easily derived as follows. By dividing both sides of equation (48c) by τ_ν one obtains

$$\frac{\tau_\nu}{\tau_{\nu-1}} = \frac{2\,M_{\nu-1}}{M_{\nu-1} + M_\nu} \; \frac{e^{i\frac{\beta\nu}{2}}}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \, . \qquad (53)$$

Now,

$$\frac{\tau_N}{\tau_o} = \frac{\tau_1}{\tau_o} \; \frac{\tau_2}{\tau_1} \cdots\cdots\cdots \frac{\tau_N}{\tau_{N-1}} = \prod_{\nu=1}^{N} \frac{2\,M_{\nu-1}}{M_{\nu-1} + M_\nu} \; \frac{e^{i\frac{\beta\nu}{2}}}{1 + W_\nu \, \rho_\nu \, e^{i\frac{\beta\nu}{2}}} \qquad (53a)$$

wherein $\prod$ denotes a product. Furthermore,

$$\frac{\tau_{N+1}}{\tau_N} = \frac{2\,M_N}{M_N + M_{N+1}} \, , \qquad (53b)$$

the Fresnel coefficient of transmission of the last interface. Therefore

$$\frac{\tau_{N+1}}{\tau_o} = \prod_{\nu=1}^{N+1} \frac{2\,M_{\nu-1}}{M_{\nu-1} + M_\nu} \; \prod_{\nu=1}^{N} \frac{e^{\frac{i\beta\nu}{2}}}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \, , \qquad (54)$$

where τ_{N+1}/τ_o is the complex transmittance from the left hand side of the first interface to the right hand side of the last interface. We observe that the complex transmittance is not merely the product of the Fresnel coefficients of transmission of the N + 1 interfaces and of a factor that includes the optical path through the layers. Consider, for example, the case in which all $K_\nu = 0$ so that $M_\nu = n_\nu$ and $\beta_\nu = \frac{4\pi}{\lambda} n_\nu \, d_\nu$. From equation (54)

$$\frac{\tau_{N+1}}{\tau_o} = e^{i\frac{2\pi}{\lambda}\sum_{\nu=1}^{N} n_\nu \, d_\nu} \; \prod_{\nu=1}^{N+1} \frac{2\,n_{\nu-1}}{n_{\nu-1} + n_\nu} \; \prod_{\nu=1}^{N} \frac{1}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \qquad (54a)$$

The quantity $\sum_{\nu=1}^{N} n_\nu \, d_\nu$ is the optical path through the layer system. The product from $\nu = 1$ to $\nu = N + 1$ is the product of the Fresnel coefficients of transmission of the interfaces. The product from $\nu = 1$ to $\nu = N$ is due to interreflections within the multilayer system. Since this product may not be real, the phase change introduced into the wave as it traverses the multilayer is not in general equal to the optical path through the layer. In designing a lens system of high optical quality, it finally becomes necessary to consider the phase changes introduced by transmission through the various layers or multilayers as the number of coated surfaces is increased. Consequently, equation (54) is of interest to both the optical designer and the designer of thin films. We shall see that similar equations hold for oblique incidence.

21.2.8.8 For reasons discussed at the end of Section 21.2.5, $|\rho_o|^2 = |r_o/\tau_o|^2$ is always energy reflectance of the multilayer. Because the incidence is normal, $i_o = i_{N+1} = 0$. Therefore, as in equation (38),

$$\text{Energy transmittance} = \frac{(n_a)_{N+1}}{n_o}\left|\frac{\tau_{N+1}}{\tau_o}\right|^2 \tag{55}$$

in which τ_{N+1}/τ_o is given by equation (54). Because $p_o = 0$ at normal incidence, a result similar to equation (31) gives $(n_a)_{N+1} = R_e(M_{N+1}) = n_{N+1}$. Consequently,

$$\text{Energy transmittance} = \frac{n_{N+1}}{n_o}\left|\frac{\tau_{N+1}}{\tau_o}\right|^2 \tag{55a}$$

at normal incidence when the initial medium is non-absorbing. When the initial and final media are identical and non-absorbing, the energy transmittance of the multilayer is simply $|\tau_{N+1}/\tau_o|^2$.

21.2.9 Oblique incidence upon multilayers; the electric vector perpendicular to the plane of incidence.

21.2.9.1 The theory of Section 21.2.8 applies again with minor changes. Let β_ν be written in the more general form

$$\beta_\nu = \frac{4\pi}{\lambda} M_\nu d_\nu , \tag{56}$$

a result that reduces to equation (47) when $p_o = 0$ (normal incidence). Then,

$$\rho_{\nu-1} = \frac{\rho_\nu e^{i\beta\nu} + W_\nu}{1 + W_\nu \rho_\nu e^{i\beta\nu}} \tag{57}$$

in which W_ν and M_ν are defined by equations (20) and (19). Because ρ_N is given by equation (51), one can compute $\rho_{N-1}, \rho_{N-2} \cdots\cdots\cdots \rho_o$ consecutively from equation (57) to obtain the complex reflectance ρ_o of the multilayer at the left hand boundary of the first interface, $z = 0$, when the electric vector is perpendicular to the plane of incidence. Furthermore, the complex transmittance, τ_{N+1}/τ_o, from the left hand side of the first interface to the right hand side of the last interface, Figure 21.6, is given again by equation (54). The quantity, $|\rho_o|^2$, is energy reflectance of the entire multilayer. The energy transmittance is given by a result similar to that of equation (38), specifically:

$$\text{Energy transmittance} = \frac{(n_a)_{N+1} \cos i_{N+1}}{n_o \cos i_o}\left|\frac{\tau_{N+1}}{\tau_o}\right|^2 , \tag{58}$$

in which

$$(n_a)_{N+1} = \left[n_o^2 \; p_o^2 + R_e^2 (M_{N+1})\right]^{1/2} \tag{59}$$

and

$$(n_a)_{N+1} \sin i_{N+1} = n_o \, p_o , \tag{60}$$

wherein the wave normals (rays) make the angles i_o and i_{N+1} with Z, in the first and last medium, respectively. See Figure 21.6. It has been assumed that the medium of incidence is non-absorbing in writing equation (58). If the first and last media are identical and non-absorbing, the energy transmittance of the multilayer is given by $|\tau_{N+1}/\tau_o|^2$.

21.2.9.2 For some purposes, it is desirable to find the directions of the wave normals in the various layers. When $U = (0, U_y, 0)$ and the time factor is $e^{-i\omega t}$, the wave equation (4) is satisfied in the ν^{th} medium or layer by a transmitted wave of the form

$$U_y = \tau_\nu \, e^{ikm_o p_o x} \; e^{-ikM_\nu (z - L_\nu)} \tag{61}$$

provided that M_ν obeys equation (19) where τ_ν specifies the amplitude and phase of the transmitted E-vector at point $x = 0$ in the interface $z = L_\nu$. Equation (61) may be written in the form

$$U_y = \tau_\nu \, e^{-ik M_\nu L_\nu} \; e^{-k(n_o K_o p_o x + I_m (M_\nu) z)} \; e^{ik(n_o p_o x + R_e (M_\nu) z)} \tag{61a}$$

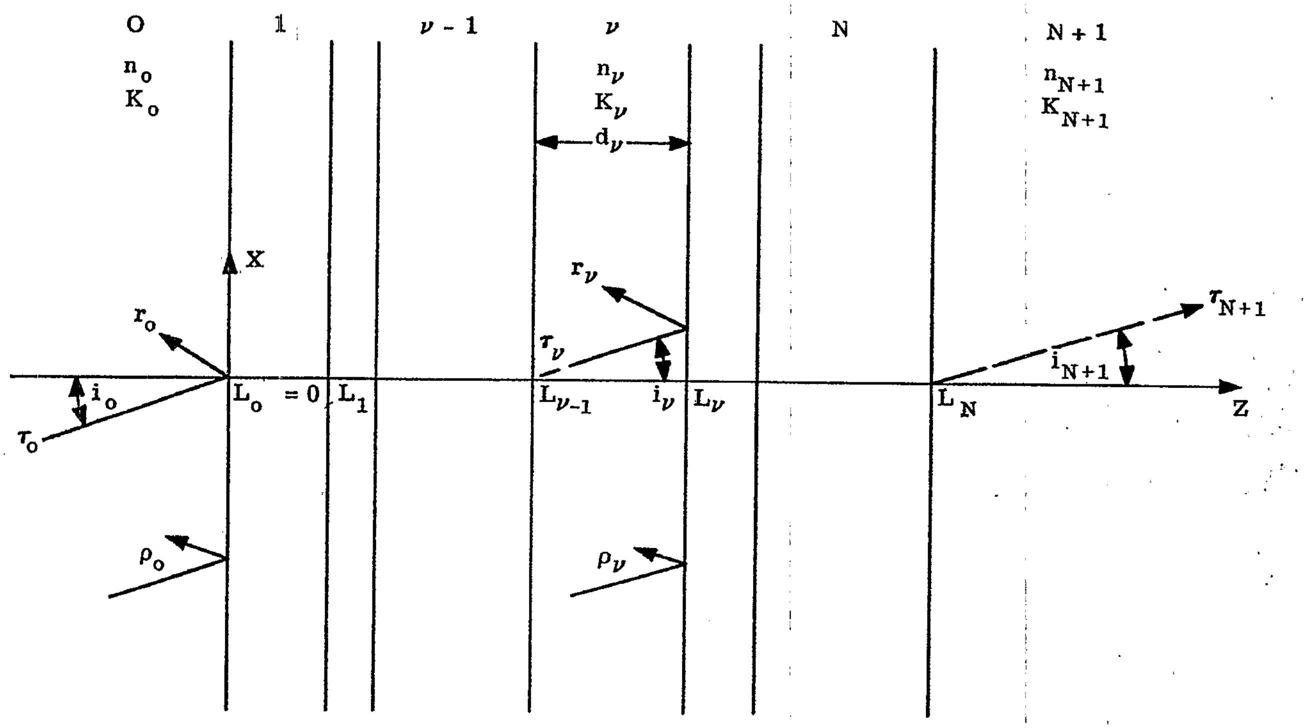


Figure 21. 6- Notation with respect to a multilayer of N films when the incidence is oblique and the E-vector is perpendicular to the plane of incidence. The angles i_ν refer to the directions of the wave normals.

Every equiphase surface in any layer ν is therefore a surface

$$n_o p_o x + R_e (M_\nu) z = \text{constant.} \tag{62}$$

If, then, the normals to the equiphase surface make the angle i_ν with Z, it follows directly* from equation (62) that

$$\sin i_\nu = n_o p_o / \left[n_o^2 p_o^2 + R_e^2 (M_\nu) \right]^{1/2} ,$$

or that

$$(n_a)_\nu \sin i_\nu = n_o \sin i_o , \tag{63}$$

wherein

$$(n_a)_\nu \equiv \left[n_o^2 p_o^2 + R_e^2 (M_\nu) \right]^{1/2} \tag{64}$$

is the actual refractive index of the νth layer. Hence the generalized Snell's Law described by equations (63) and (64) holds with respect to all of the media of the system. In particular, equations (59) and (60) are special cases of equations (64) and (63).

21. 2. 9. 3 When required, the electromagnetic field in any layer or medium can be computed as follows. In each medium

$$(E_\nu)_{\text{transmitted}} = (0, 1, 0) \cdot \tau_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu (z - L_\nu)} , \tag{65}$$

wherein $E_{\text{transmitted}}$ is the wave propagated to the right, Figure 21. 6. The corresponding H-vector is now computed from equations (27) and stated with the aid of equation (1). With respect to the wave propagated to the left in each medium

---

* See discussions of the normal form of the equation of a straight line in textbooks on analytic geometry.

$$(E_\nu)_{\text{reflected}} = (0,\ 1,\ 0)\, r_\nu\, e^{-i\omega t}\, e^{ikm_o p_o x}\, e^{-ikM_\nu (z - L_\nu)} . \qquad (66)$$

The corresponding "reflected" H-vector is computed again from equation (27) and stated with the aid of equation (1). When all ρ_ν's have been determined, each τ_ν can be computed from $\tau_{\nu-1}$ by means of equation (53). Of course, τ_o must be given or assigned. Next, all r_ν's can be calculated from the known ratios $\rho_\nu = r_\nu/\tau_\nu$ and the known values of τ_ν. The theory is therefore a complete theory for the case in which the electric vector is perpendicular to the plane X, Z of incidence, Figure 21. 6.

21. 2. 10 Oblique incidence upon multilayers; the magnetic vector perpendicular to the plane X, Z of incidence.

21. 2. 10. 1 As can be expected, the theory of oblique incidence upon multilayers with the magnetic vector perpendicular to the plane X, Z of incidence differs from the case in which the electric vector is perpendicular to the plane of incidence only in the Fresnel coefficients that become involved. Let

$$\gamma_\nu = R_\nu/T_\nu , \qquad (67)$$

where T_ν is a complex number that specifies the amplitude and phase of the H-vector propagated to the right, Figure 21. 7, at $x = 0$ in the right hand boundary of the ν^{th} medium, and where R_ν specifies the amplitude and phase of the H-vector propagated to the left at point $x = 0$ in the right hand boundary of the ν^{th} medium, for the range of ν from $\nu = 0$ to $\nu = N$. $R_{N+1} = 0$ because no reflected wave exists in the last medium. T_{N+1} specifies the amplitude and phase of the transmitted H-vector at the left hand boundary of the $(N + 1)^{\text{th}}$ medium. Then

$$\gamma_{\nu-1} = \frac{\gamma_\nu\, e^{i\beta_\nu} + F_\nu}{1 + F_\nu\, \gamma_\nu\, e^{i\beta_\nu}} , \qquad (68)$$

in which β_ν is given by equation (56) and the Fresnel coefficients of reflection F_ν are given by equation (21).

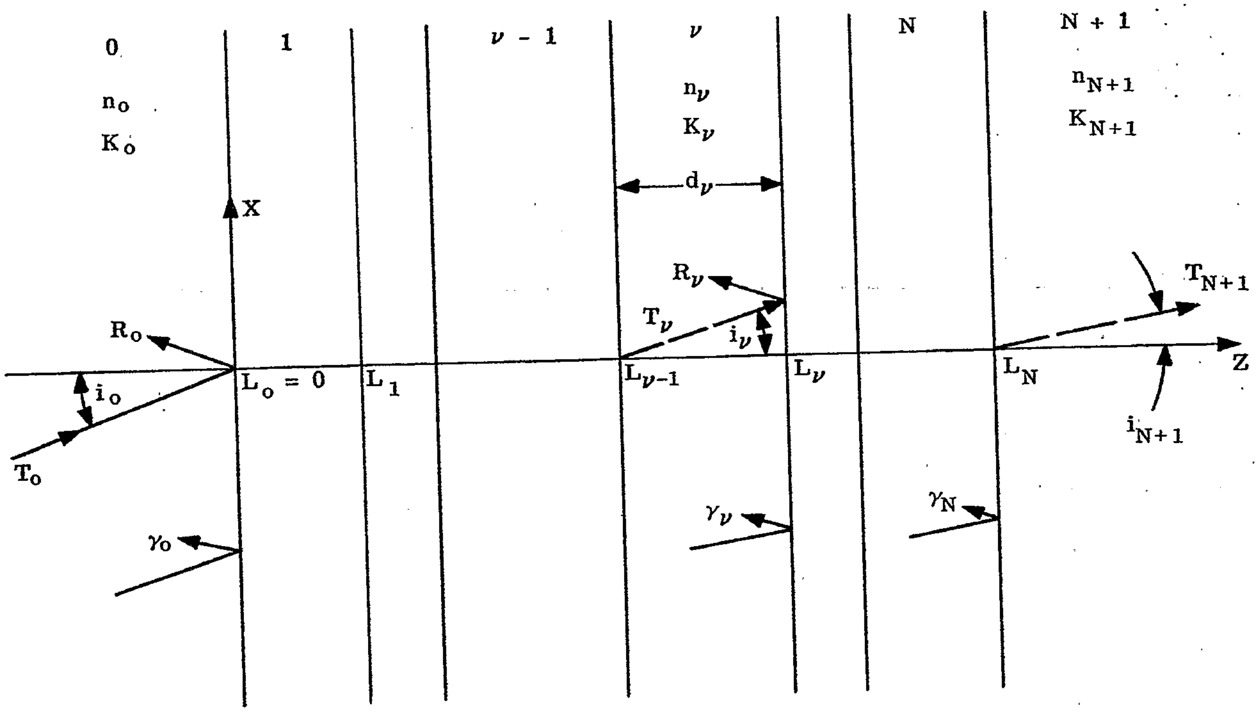


Figure 21. 7- Notation for the case in which the magnetic vector is polarized to vibrate perpendicular to the plane X, Z of incidence. The wave normals have the same angles i_ν with Z as in Figure 21. 5.

Since

$$\gamma_N = F_{N+1}, \qquad (69)$$

the recursion formula (68) enables one to compute consecutively all complex reflectances, γ_ν from γ_N down to γ_o. These complex reflectances, γ_ν, are defined at the right hand boundary of the νth medium, as in the case of the complex reflectances ρ_ν.

21.2.10.2 The complex transmittance, T_{N+1}/T_o, from the left hand boundary of the first interface, $z = 0$, to the right hand boundary of the last interface, $z = L_N$, is given by

$$\frac{T_{N+1}}{T_o} = \prod_{\nu=1}^{N+1} \frac{2\, M_{\nu-1}\, m_\nu^2}{M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2} \prod_{\nu=1}^{N} \frac{e^{\frac{i\beta_\nu}{2}}}{1 + F_\nu\, \gamma_\nu\, e^{i\beta_\nu}}, \qquad (70)$$

in which $2M_{\nu-1}\, m_\nu^2 / \left[M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2\right]$ is the Fresnel coefficient of transmission in the forward direction through the interface between the $(\nu - 1)$ th and νth medium. Equation (70) should be compared with equation (54).

21.2.10.3 The energy reflectance is always given in this state of polarization by $|\gamma_o|^2$. Calculation of the energy transmittance involves difficulties, as discussed at the end of Section 21.2.6, unless the time-averaged Poynting vector is practically parallel to the wave normals. If K_{N+1} is so small that the Poynting vector is sensibly parallel to the wave normals in the last medium, then, as in equation (45b),

$$\text{Energy transmittance} = \frac{n_{N+1}^2\, n_o (n_a)_{N+1} \cos i_{N+1}}{|m_{N+1}|^4 \cos i_o} \left|\frac{T_{N+1}}{T_o}\right|^2, \qquad (71)$$

in which the medium of incidence is assumed to be non-absorbing. If, also, the last medium is non-absorbing,

$$\text{Energy transmittance} = \frac{n_o \cos i_{N+1}}{n_1 \cos i_o} \left|\frac{T_{N+1}}{T_o}\right|^2. \qquad (71a)$$

21.2.10.4 The electromagnetic field in the νth medium is determined as follows. In each medium

$$(H_\nu)_{\text{transmitted}} = (0,\ 1,\ 0)\ T_\nu\ e^{-i\omega t}\ e^{ikm_o p_o x}\ e^{ikM_\nu (z - L_\nu)}, \qquad (72)$$

for the wave propagated to the right, Figure 21.7. Compare with equation (65). The corresponding E-vector is determined by equation (41) and stated with the aid of equation (1). With respect to the wave propagated to the left in the νth medium,

$$(H_\nu)_{\text{reflected}} = (0,\ 1,\ 0)\ R_\nu\ e^{-i\omega t}\ e^{ikm_o p_o x}\ e^{-ikM_\nu (z - L_\nu)}. \qquad (73)$$

The corresponding reflected E-vector can be computed from equation (41) and stated with the aid of equation (1). When the method of 21.2.8 is applied to the case in which the H-vector is perpendicular to the X, Z plane, one finds instead of equation (53), that

$$\frac{T_\nu}{T_{\nu-1}} = \frac{2\, M_{\nu-1}\, m_\nu^2}{M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2}\ \frac{e^{i\frac{\beta_\nu}{2}}}{1 + F_\nu\, \gamma_\nu\, e^{i\beta_\nu}}. \qquad (74)$$

If all values of γ_ν have been computed from equation (68), every T_ν from T_1 to T_N can be computed from equation (74). Since $\gamma_\nu = R_\nu/T_\nu$, every R_ν can be computed so that the coefficients R_ν and T_ν of the electromagnetic fields become known.

21.2.10.5 Comparison of equations (65) and (66) with equations (72) and (73) shows that the corresponding exponential factors are alike. Consequently, the wave normals are refracted from one layer into the next so as to make the same angle, i_ν, with Z whether the E-vector or the H-vector is perpendicular to the plane of incidence.

21.2.11 An approximate method of computation based upon the complex reflectances.

21.2.11.1 Comparison of equations (50), (57) and (68) shows that the recursion formula (50) for normal incidence can be used as the prototype in dealing with the complex reflectances. It is necessary only to enter the appropriate set of Fresnel coefficients, W_ν or F_ν, and to insert the value of $p_o = \sin i_o$ into β_ν as defined by equations (56) and (19). The following approximation simplifies and expedites the determination of ρ_o. The approximation becomes excellent for films that are intended to exhibit low reflectance and have relatively small Fresnel coefficients, W_ν or F_ν. Inspection of equation (50) reveals at once that when both W_ν and

ρ_ν are small, $\rho_{\nu-1}$ should be given with good approximation by

$$\rho_{\nu-1} = \rho_\nu \, e^{i\beta\nu} + W_\nu . \tag{75}$$

Thus

$$\rho_N = W_{N+1} ,$$
$$\rho_{N-1} = W_{N+1} \, e^{i\beta_N} + W_N ,$$
$$\rho_{N-2} = W_{N+1} \, e^{i(\beta_N + \beta_{N-1})} + W_N \, e^{i\beta_{N-1}} + W_{N-1} ,$$

and therefore,

$$\rho_o = \sum_{\nu=1}^{N+1} W_\nu \exp i \sum_{\mu=1}^{\nu-1} \beta\mu . \tag{76}$$

When the sum (76) is computed as the complex number

$$\rho_o = R_e (\rho_o) + i \, I_m (\rho_o) , \tag{76a}$$
$$|\rho_o|^2 = R_e^2 (\rho_o) + I_m^2 (\rho_o) . \tag{76b}$$

21.2.11.2 This approximate method for computing ρ_o is used mainly during rapid exploration for likely multilayer systems that do not contain absorbing layers and that are intended to produce low reflectance. When absorption is absent, one may compute $|\rho_o|^2$ directly from the sum

$$|\rho_o|^2 = \sum_{\nu=1}^{N+1} W_\nu + 2 \sum_{\mu=1}^{N} \sum_{\nu=2}^{N+1} W\mu \, W_\nu \cos(\beta_\mu + \beta_{\mu+1} + \cdots\cdots\cdots \beta_{\nu-1}), \tag{76c}$$

in which $\mu < \nu$. Comparison of $|\rho_o|_a^2$ computed from the approximate equation (76c) with the values $|\rho_o|^2$ computed from equation (50) is made in Figure 21.8 for the case of a low reflecting trilayer. The Fresnel coefficients W_2 and W_3 are quite large numerically.

21.2.11.3 The approximate method of equation (76) is the algebraic equivalent of the graphical polygon method used in early calculations on mono and bilayers by C. H. Cartwright (2) and others.

21.2.12 Method of admittances; E-vector perpendicular to the plane of incidence.

21.2.12.1 Because the theories of multilayers and transmission lines are similar, many investigators prefer to treat a multilayer as a transmission line. One of the earlier publications dealing with multilayers in terms of the admittances of transmission lines is due to B. Salzberg (3). It will be one aim of the following presentation to unify and to show the relationships that exist between the optical method of the reflectances and the electrical method of the admittances. These two methods are equivalent and complementary. Each method possesses some advantages over the other in dealing with thin films.

21.2.12.2 From equations (65) and (66)

$$(E_{y,\nu})_{\text{transmitted}} = \tau_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{ikM_\nu (z - L_\nu)} , \tag{77}$$

and

$$(E_{y,\nu})_{\text{reflected}} = r_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{-ikM_\nu (z - L_\nu)} , \tag{77a}$$

wherein the subscript ν refers to the wave in the νth medium or layer. Correspondingly from equations (27) and (1),

$$(H_{x,\nu})_{\text{transmitted}} = - \tau_\nu M_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{ikM_\nu (z - L_\nu)} , \tag{77b}$$

and

$$(H_{x,\nu})_{\text{reflected}} = r_\nu M_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{-ikM_\nu (z - L_\nu)} . \tag{77c}$$

---

(2) C. H. Cartwright et al, U. S. Patent 2, 281, 474.

(3) Bernard Salzberg, J. Opt. Soc. Amer., 40, 465-470 (1950).

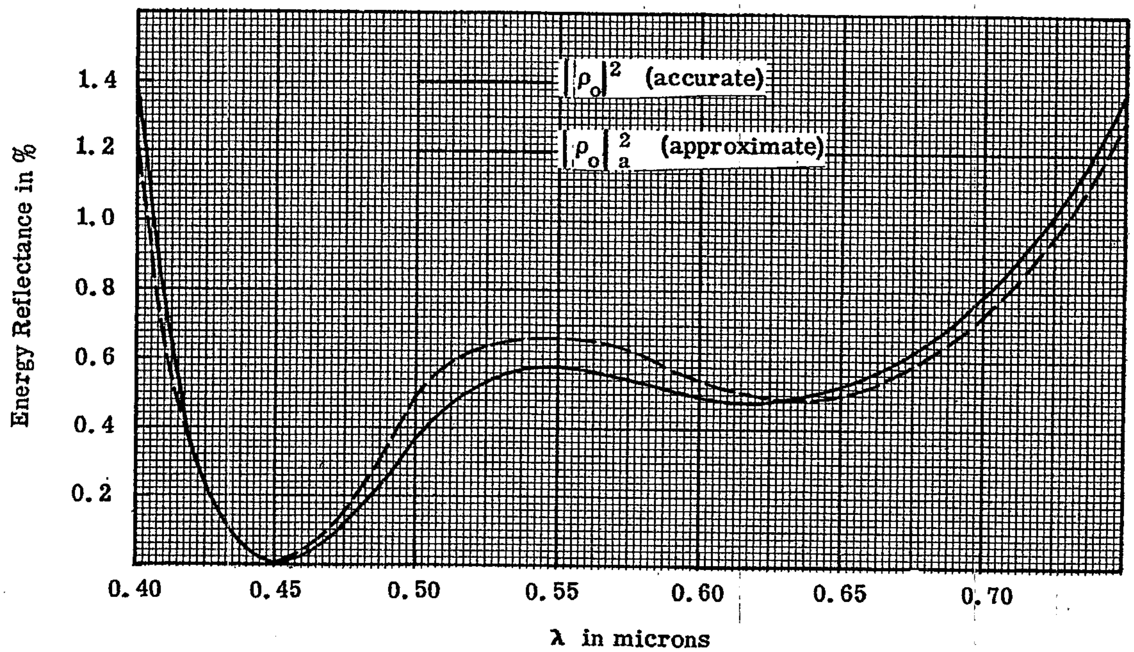


Figure 21. 8- Comparison of the computed energy reflectances $|\rho_o|^2$ and $|\rho_o|_a^2$ for a tri-layer in which $W_1 = -0.15966$; $W_2 = -0.28866$; $W_3 = 0.18765$ and $W_4 = 0.05882$. At the wavelength $\lambda_o = 0.55$ microns, $\beta_1 = 235.8^o$; $\beta_2 = 32^o$ and $\beta_3 = 360^o$.

The admittance Y_ν for the case in which the electric vector is perpendicular to the plane of incidence is defined so that

$$Y_\nu = \frac{(H_{x,\nu})_{\text{transmitted}} + (H_{x,\nu})_{\text{reflected}}}{(E_{y,\nu})_{\text{transmitted}} + (E_{y,\nu})_{\text{reflected}}}, \qquad (78)$$

evaluated at the left hand boundary, $z = L_{\nu-1}$, of the νth medium or layer. Since $L_{\nu-1} - L_\nu = -d_\nu$, equations (77) through (78) give almost directly the result

$$Y_\nu = M_\nu \frac{-\tau_\nu e^{-i\frac{\beta_\nu}{2}} + r_\nu e^{i\frac{\beta_\nu}{2}}}{\tau_\nu e^{-i\frac{\beta_\nu}{2}} + r_\nu e^{i\frac{\beta_\nu}{2}}}, \qquad (78a)$$

or

$$Y_\nu = M_\nu \frac{-1 + \rho_\nu e^{i\beta_\nu}}{1 + \rho_\nu e^{i\beta_\nu}}, \qquad (78b)$$

because $\rho_\nu = r_\nu / \tau_\nu$. Upon solving equation (78b) for $\rho_\nu e^{i\beta_\nu}$, one obtains

$$\rho_\nu e^{i\beta_\nu} = \frac{M_\nu + Y_\nu}{M_\nu - Y_\nu}. \qquad (78c)$$

Hence the admittances Y_ν can be computed from the reflectances ρ_ν and vice versa.

From equation (57),

$$\rho_o = \frac{\rho_1 e^{i\beta_1} + W_1}{1 + W_1 \rho_1 e^{i\beta_1}}. \qquad (79)$$

Upon eliminating $\rho_1 e^{i\beta_1}$ with the aid of equation (78c) and making use of the identity

$$M_\nu \frac{1 + W_\nu}{1 - W_\nu} = M_{\nu-1}, \qquad (80)$$

one finds that the complex reflectance, ρ_o, of the multilayer, and the admittance, Y_1, are connected by the equation

$$\rho_o = \frac{M_o + Y_1}{M_o - Y_1} . \tag{79a}$$

From equation (78b),

$$Y_{\nu-1} = M_{\nu-1} \frac{\rho_{\nu-1} e^{i\beta_{\nu-1}} - 1}{\rho_{\nu-1} e^{i\beta_{\nu-1}} + 1} . \tag{81}$$

Just as equation (79a) is obtained from equation (79),

$$\rho_{\nu-1} = \frac{M_{\nu-1} + Y_\nu}{M_{\nu-1} - Y_\nu} . \tag{82}$$

By eliminating $\rho_{\nu-1}$ from equation (81) with the aid of equation (82) and utilizing the identity

$$\frac{e^{-ix} - 1}{e^{-ix} + 1} = - \tanh \frac{ix}{2} = - i \tan \frac{x}{2} ,$$

one obtains without difficulty the recursion formula for $Y_{\nu-1}$ in the form

$$Y_{\nu-1} = M_{\nu-1} \frac{Y_\nu - i M_{\nu-1} \tan(\beta_{\nu-1}/2)}{M_{\nu-1} - i Y_\nu \tan(\beta_{\nu-1}/2)} . \tag{81a}$$

Because $\rho_{N+1} = 0$ (no reflected wave in the last medium), it follows from equation (78c) as a boundary condition that

$$Y_{N+1} = - M_{N+1} . \tag{81b}$$

Hence the recursion formula (81a) enables one to compute all Y_ν's from Y_N down to Y_1. With Y_1 thus determined, equation (79a) can be applied to find the complex reflectance, ρ_o, of the multilayer.

21.2.12.3 The complex transmittance, τ_{N+1}/τ_o, of the multilayer is now given awkwardly by equation (54). From equation (78c),

$$1 + W_\nu \rho_\nu e^{i\beta\nu} = 1 + W_\nu \left(\frac{M_\nu + Y_\nu}{M_\nu - Y_\nu} \right) = \frac{M_\nu (1 + W_\nu) - Y_\nu (1 - W_\nu)}{M_\nu - Y_\nu} ;$$

$$= (1 - W_\nu) \frac{M_{\nu-1} - Y_\nu}{M_\nu - Y_\nu} . \tag{82a}$$

Since

$$1 - W_\nu = 2 M_\nu / (M_{\nu-1} + M_\nu) , \tag{82b}$$

$$1 + W_\nu \rho_\nu e^{i\beta\nu} = \frac{2 M_\nu}{M_{\nu-1} + M_\nu} \frac{M_{\nu-1} - Y_\nu}{M_\nu - Y_\nu} . \tag{82c}$$

Let $1 + W_\nu \rho_\nu e^{i\beta\nu}$ be eliminated from equation (54) with the aid of equation (82c). Then

$$\frac{\tau_{N+1}}{\tau_o} = \frac{2 M_N}{M_N + M_{N+1}} \exp i \left(\sum_{\nu=1}^{N} \frac{\beta_\nu}{2} \right) \prod_{\nu=1}^{N} \left[\left(\frac{M_{\nu-1}}{M_\nu} \right) \frac{M_\nu - Y_\nu}{M_{\nu-1} - Y_\nu} \right] , \tag{83}$$

a result that should be compared with equation (54).

21.2.12.4 In summary, when the electric vector is perpendicular to the plane X, Z of incidence, the complex reflectance ρ_o and the complex transmittance τ_{N+1}/τ_o of the multilayer can be computed from equations (79a) and (83) in which all the admittances, Y_ν, from Y_N down to Y_1 are determined by the recursion formula (81a). The admittance, Y_ν, at the point of entry into the νth layer is defined by equations (78) or (78a). β_ν and M_ν are defined by equations (56) and (19), respectively. The results for normal incidence are obtained by setting $p_o = 0$ in equation (19), i.e. by setting $M_\nu = m_\nu = n_\nu (1 + i K_\nu)$.

21. 2. 13 Method of admittances; H-vector perpendicular to the plane of incidence.

21. 2. 13. 1 When the H-vector is polarized to vibrate at right angles to the plane X, Z of incidence, equations (72) and (73) show that

$$(H_{y,\nu})_{\text{transmitted}} = T_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu(z - L_\nu)} , \qquad (84)$$

and

$$(H_{y,\nu})_{\text{reflected}} = R_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{-ikM_\nu(z - L_\nu)} . \qquad (84a)$$

Correspondingly, from equations (41) and (1),

$$(E_{x,\,\nu})_{\text{transmitted}} = \frac{M_\nu}{m_\nu^2} \; T_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu(z - L_\nu)} , \qquad (84b)$$

and

$$(E_{x,\,\nu})_{\text{reflected}} = \frac{-M_\nu}{m_\nu^2} \; R_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{-ikM_\nu(z - L_\nu)} . \qquad (84c)$$

The admittances are designated by y_ν to distinguish them from the admittances Y_ν of equation (78) and are defined by the equation

$$y_\nu = \frac{(H_{y,\,\nu})_{\text{transmitted}} + (H_{y,\,\nu})_{\text{reflected}}}{(E_{x,\,\nu})_{\text{transmitted}} + (E_{x,\,\nu})_{\text{reflected}}} , \qquad (85)$$

evaluated at the point of entry, $z = L_{\nu-1}$, of the ν^{th} layer. Substitution of equation (84) into equation (85) yields the result

$$y_\nu = \frac{m_\nu^2}{M_\nu} \; \frac{1 + \gamma_\nu \; e^{i\beta_\nu}}{1 - \gamma_\nu \; e^{i\beta_\nu}} , \qquad (85a)$$

in which $\gamma_\nu = R_\nu / T_\nu$ as in equation (67). It follows from equation (85a) that

$$\gamma_\nu \; e^{i\beta_\nu} = \frac{y_\nu - m_\nu^2/M_\nu}{y_\nu + m_\nu^2/M_\nu} . \qquad (85b)$$

21. 2. 13. 2 Corresponding to the identity (80), one finds from the definition of F_ν, equation (21), that

$$\frac{m_\nu^2}{M_\nu} \; \frac{1 - F_\nu}{1 + F_\nu} = \frac{m_{\nu-1}^2}{M_{\nu-1}} . \qquad (86)$$

Let $\gamma_\nu \; e^{i\beta_\nu}$ be eliminated from equation (68) with the aid of equation (85b). By utilizing the identity (86) in the result thus obtained, one finds straightforwardly that

$$\gamma_{\nu-1} = \frac{y_\nu - m_{\nu-1}^2/M_{\nu-1}}{y_\nu + m_{\nu-1}^2/M_{\nu-1}} . \qquad (87)$$

In particular,

$$\gamma_o = \frac{y_1 - m_o^2/M_o}{y_1 + m_o^2/M_o} , \qquad (87a)$$

where γ_o is the complex reflectance of the entire multilayer evaluated at the first interface, $z = 0$, Figure 21. 7.

21. 2. 13. 3 From equation (85a),

$$y_{\nu-1} = \frac{m_{\nu-1}^2}{M_{\nu-1}} \; \frac{1 + \gamma_{\nu-1} \; e^{i\beta_{\nu-1}}}{1 - \gamma_{\nu-1} \; e^{i\beta_{\nu-1}}} \qquad (88)$$

Let $\gamma_{\nu-1}$ be eliminated from equation (88) with the aid of equation (87). As in the steps leading to equation

(81a), one finds that

$$y_{\nu-1} = \frac{m^2_{\nu-1}}{M_{\nu-1}} \left[\frac{y_\nu - i \dfrac{m^2_{\nu-1}}{M_{\nu-1}} \tan(\beta_{\nu-1}/2)}{\dfrac{m^2_{\nu-1}}{M_{\nu-1}} - i\, y_\nu \tan(\beta_{\nu-1}/2)} \right] . \qquad (88a)$$

This recursion formula is similar to equation (81a). It is necessary only to replace $M_{\nu-1}$ by the ratio $m^2_{\nu-1}/M_{\nu-1}$. Since $\gamma_{N+1} = 0$, equation (85b) shows that a boundary condition on y_ν is

$$y_{N+1} = m^2_{N+1}/M_{N+1} . \qquad (88b)$$

Equations (88a) and (88b) permit all values of y_ν to be computed from y_N down to y_1 from the optical properties of the multilayer system. At this point, the solution for the complex reflectance, γ_o, of the multilayer can be completed from equation (87a).

21.2.13.4 It remains to develop a more suitable formula to replace equation (70) for the complex transmittance T_{N+1}/T_o of the multilayer. From equation (85b),

$$1 + F_\nu \gamma_\nu e^{i\beta_\nu} = 1 + F_\nu \left(\frac{y_\nu - m_\nu^2/M_\nu}{y_\nu + m_\nu^2/M_\nu} \right) = \frac{y_\nu (1 + F_\nu) + \dfrac{m_\nu^2}{M_\nu} (1 - F_\nu)}{y_\nu + m_\nu^2/M_\nu}$$

$$= (1 + F_\nu) \frac{y_\nu + m^2_{\nu-1}/M_{\nu-1}}{y_\nu + m_\nu^2/M_\nu} ; \qquad (89)$$

with

$$1 + F_\nu = \frac{2 m_\nu^2 M_{\nu-1}}{m_\nu^2 M_{\nu-1} + m^2_{\nu-1} M_\nu} . \qquad (89a)$$

Let $1 + F_\nu \gamma_\nu e^{i\beta_\nu}$ be eliminated from equation (70) with the aid of equations (89) and (89a). Then

$$\frac{T_{N+1}}{T_o} = \frac{2 M_N m^2_{N+1}}{M_N m^2_{N+1} + M_{N+1} m_N^2} \exp\left(i \sum_{\nu=1}^{N} \frac{\beta_\nu}{2} \right) \prod_{\nu=1}^{N} \frac{y_\nu + m_\nu^2/M_\nu}{y_\nu + m^2_{\nu-1}/M_\nu} . \qquad (89b)$$

21.2.13.5 In summary: When the magnetic vector vibrates at right angles to the plane X, Z of incidence, the complex reflectance γ_o and the complex transmittance T_{N+1}/T_o of the multilayer are determined by equations (87a) and (89b) when the admittances, y_ν, from y_N to y_1 have been computed from the recursion formula (88a). The admittances, y_ν, are defined by equations (85) or (85a), and are different from the complementary admittances, Y_ν, for the state of polarization in which the electric vector is perpendicular to the plane X, Z of incidence. β_ν and M_ν are defined by equations (56) and (19). When desired, the admittances, y_ν, can be computed from the complex reflectances, γ_ν, by means of equation (85a); or the complex reflectances, γ_ν, can be computed from the admittances, y_ν, by means of equation (85b).

21.2.14 Absentee layers.

21.2.14.1 It can be shown easily from the method of admittances that non-absorbing layers behave as if they were absent at wavelengths λ for which

$$n_\nu d_\nu \cos i_\nu = \mu \frac{\lambda}{2} ; \quad \mu = 1, 2, 3, 4, \text{etc.} \qquad (90)$$

We shall call such layers absentee layers. For example, at normal incidence the so called half-wave layer, the case $\mu = 1$ and $i_\nu = 0$ in equation (90), is an absentee layer. Condition (90) is satisfied when $\beta_\nu = \mu\, 2\pi$, i.e. when $\tan(\beta_\nu/2) = 0$ in equations (81a) and (88a). Consequently, $Y_{\nu-1} = Y_\nu$ and $y_{\nu-1} = y_\nu$ when equation (90) is satisfied. This means that the νth layer does not affect the admittance at the point of entry of the $(\nu-1)$th layer whether the E-vector or the H-vector vibrates at right angles to the plane X, Z of incidence. The νth layer does not influence the reflectance or the transmittance of the multilayer at any wavelength for which equation (90) is satisfied. In fact, the layer behaves as if $\mu = 0$, i.e. as if the thickness d_ν of the layer were zero.

21.2.14.2 Consider the behavior, with wavelength λ_μ or with β_ν, of a single, homogeneous, non-absorbing film deposited on any substrate. Because this film is an absentee layer at wavelengths λ_μ for which equation (90) is satisfied, it follows, for example, that the energy reflectances $|\rho_o|^2$ of the coated and uncoated surface should be alike at the wavelengths λ_μ. Actually, these reflectances are not quite alike at λ_μ because the film may absorb, may not be homogeneous, or may scatter appreciably.

21.2.14.3 Absentee layers are often introduced in multilayers for the purpose of altering the behavior of the multilayer at wavelengths λ that do not satisfy equation (90).

21.2.15 The Q-Method.

21.2.15.1 Comparison of the recursion formulae (57), (68), (81a) and (88a) shows that they are all awkward for the purposes of computation. With respect to equation (57), for example, $\rho_{\nu-1}$ is not linear in ρ_ν. This lack of linearity applies to the reflectances and the admittances alike. The reflectances enjoy a small advantage in that approximate expressions such as equation (75) exhibit linearity. It is one of the purposes of the Q-method(4) to circumvent the lack of linearity in the recursion formulae for the reflectances and the admittances.

21.2.15.2 The recurrence formulae connecting successive interfacial reflectances or admittances are of the form

$$A_{\nu-1} = \frac{a_\nu A_\nu + b_\nu}{g_\nu A_\nu + h_\nu} , \tag{91}$$

in which A_ν can represent ρ_ν, γ_ν, Y_ν or y_ν. Since the denominators of (91) are not zero, let

$$f_\nu \equiv g_\nu A_\nu + h_\nu , \tag{91a}$$

and set

$$Q_\nu \equiv \prod_{n=\nu}^{n=m+1} f_n , \tag{91b}$$

in which the upper limit, m, of the product is an integer such that

$$A_{m+1} = 0 ; \quad A_m = 0 . \tag{91c}$$

Let

$$\gamma_\nu = a_\nu / g_\nu . \tag{91d}$$

Then it can be shown that Q_ν obeys the linear recursion formula

$$Q_\nu = (\alpha_{\nu+1} g_\nu + h_\nu) Q_{\nu+1} + (b_{\nu+1} - \alpha_{\nu+1} h_{\nu+1}) g_\nu Q_{\nu+2} . \tag{91e}$$

Furthermore,

$$A_\nu = \left[\alpha_{\nu+1} Q_{\nu+1} + (b_{\nu+1} - \alpha_{\nu+1} h_{\nu+1}) Q_{\nu+2} \right] / Q_{\nu+1} . \tag{91f}$$

21.2.15.3 Consider, for example, the application of the Q-method to the determination of the complex reflectances $\rho_{\nu-1}$ of equation (57). Thus,

$$A_{\nu-1} = \rho_{\nu-1} = \frac{\rho_\nu e^{i\beta\nu} + W_\nu}{W_\nu \rho_\nu e^{i\beta\nu} + 1} . \tag{92}$$

Comparison of equations (91) and (92) shows that

$$\begin{aligned} a_\nu &= e^{i\beta\nu} ; \quad & g_\nu &= W_\nu e^{i\beta\nu} ; \\ b_\nu &= W_\nu ; \quad & h_\nu &= 1 . \end{aligned} \tag{92a}$$

Therefore,

$$\alpha_\nu = 1/W_\nu ; \tag{92b}$$

$$f_\nu = 1 + W_\nu \rho_\nu e^{i\beta\nu} . \tag{92c}$$

---

(4) H. Osterberg, J. Opt. Soc. Amer., 43, 728-732 (1953).

Because $\rho_{N+1} = 0$ in a system of multilayers containing N layers, $A_{N+1} = 0$. Hence $m = N$ in equation (91b) so that

$$Q_\nu = \prod_{n=\nu}^{N+1} f_n = \prod_{n=\nu}^{N+1} (1 + W_n \rho_n e^{i\beta n}). \tag{92d}$$

In particular,

$$Q_{N+1} = 1 ; \tag{92e}$$

$$Q_N = 1 + W_N W_{N+1} e^{i\beta_N} ; \rho_N = W_{N+1} . \tag{92f}$$

From equations (91e), (92a) and (92b),

$$Q_\nu = Q_{\nu+1} + \left[Q_{\nu+1} + (W_{\nu+1}^2 - 1) Q_{\nu+2} \right] \frac{W_\nu e^{i\beta\nu}}{W_{\nu+1}} . \tag{92g}$$

Every Q_ν from $\nu = N + 1$ down to $\nu = 1$ is determined from equations (92e) through (92g). Since $A_\nu = \rho_\nu$, equations (91f), (92a) and (92b) now yield ρ_ν in the form

$$\rho_\nu = \frac{Q_{\nu+1} + (W_{\nu+1}^2 - 1) Q_{\nu+2}}{W_{\nu+1} Q_{\nu+1}} . \tag{92h}$$

The complex reflectance, ρ_o, of the entire multilayer is therefore given by

$$\rho_o = \frac{Q_1 + (W_1^2 - 1) Q_2}{W_1 Q_1} . \tag{93}$$

21.2.15.4 With respect to the computation of the complex transmittance, τ_{N+1}/τ_o, of the multilayer, from equation (54), we note that

$$\prod_{\nu=1}^{N} \frac{1}{1 + W_\nu \rho_\nu e^{i\beta\nu}} = f_{N+1} \prod_{\nu=1}^{N+1} \frac{1}{f_\nu} = \frac{1}{Q_1} , \tag{94}$$

because $f_{N+1} = 1$. Hence the complex transmittance, τ_{N+1}/τ_o, of the multilayer is given by the comparatively simple result,

$$\frac{\tau_{N+1}}{\tau_o} = \frac{2^{N+1}}{Q_1} \exp\left(i \sum_{\nu=1}^{N} \frac{\beta_\nu}{2} \right) \prod_{\nu=1}^{N+1} \frac{M_{\nu-1}}{M_{\nu-1} + M_\nu} , \tag{95}$$

when the electric vector is perpendicular to the plane of incidence. Since β_ν is defined by equation (56),

$$\frac{\beta_\nu}{2} = \frac{2\pi d_\nu}{\lambda} \left[R_e (M_\nu) + i \ I_m (M_\nu) \right] . \tag{95a}$$

21.2.15.5 Suppose that the system has no absorption. Then $M_\nu = n_\nu \cos i_\nu$, a real number, and $\frac{\beta_\nu}{2} = \frac{2\pi}{\lambda} n_\nu d_\nu \cos i_\nu$. It will be seen from Figure 21.6 that $\frac{\beta_\nu}{2}$ is the optical path along the rays only when the incidence is normal so that $\cos i_\nu = 1$. If the incidence is normal, and if the system has no absorption, $\sum_{\nu=1}^{N} \frac{\beta_\nu}{2}$ is physically the optical path through the multilayer. But the optical path will not be equal to the phase change suffered by the wave on passing through the multilayer system unless Q_1 is real, a condition that holds only at certain wavelengths for a given multilayer.

21.2.15.6 The phase change (retardation) suffered by the wave in passing through the multilayer is equal to $\arg(\tau_{N+1}/\tau_o)$. The portion determined by $\arg(\bar{Q}_1) = \arg(1/Q_1)$ can be oscillatory with wavelength and is therefore quite dispersive. We learn that the quantity Q_1 has definite physical significance and that the Q-method is not merely a more convenient method for computing reflectance and transmittance of multilayers.

21.2.15.7 Examination of equation (92g) shows that when every β_ν is an integral multiple of the smallest value of β, each Q_ν will be a terminating exponential series of the Fourier type. Hence the powerful methods of Fourier series may be brought to bear. In the simplest case, each layer can be a quarter-wave layer at a specified wavelength. Furthermore, equation (92g) is a difference equation, consequently the methods of difference equations can often be utilized to simplify the determination of Q_1.

21.2.16 The zero condition.

21.2.16.1 One of the most important applications of single or multilayers is to reduce the energy reflectance of the coated surface. Whereas it is not always practical to attain zero reflectance, the designer of low reflecting films attempts to achieve zero reflectance at one or more wavelengths. Unfortunately, it is not possible to obtain zero reflectance over an extended range of wavelengths. Each method of attack on the design of thin films will include an analytical statement of the zero condition, i.e. the condition for obtaining zero reflectance.

21.2.16.2 With respect to the method involving the interfacial reflectances ρ_ν or γ_ν, the zero condition requires that ρ_o or $\gamma_o = 0$ as indicated in Figure 21.9. From equation (57),

$$\rho_o = \frac{\rho_1 e^{i\beta_1} + W_1}{1 + W_1 \rho_1 e^{i\beta_1}} , \qquad (96)$$

and from equation (68),

$$\gamma_o = \frac{\gamma_1 e^{i\beta_1} + F_1}{1 + F_1 \gamma_1 e^{i\beta_1}} . \qquad (97)$$

Accordingly, the zero condition assumes the form

$$\rho_1 e^{i\beta_1} = -W_1 , \qquad (98)$$

or

$$\gamma_1 e^{i\beta_1} = -F_1 , \qquad (98a)$$

depending upon whether the electric or the magnetic vector, respectively, is perpendicular to the plane of incidence. In considering normal incidence, one ordinarily chooses equation (98) and sets $i_o = 0$ in determining M_ν.

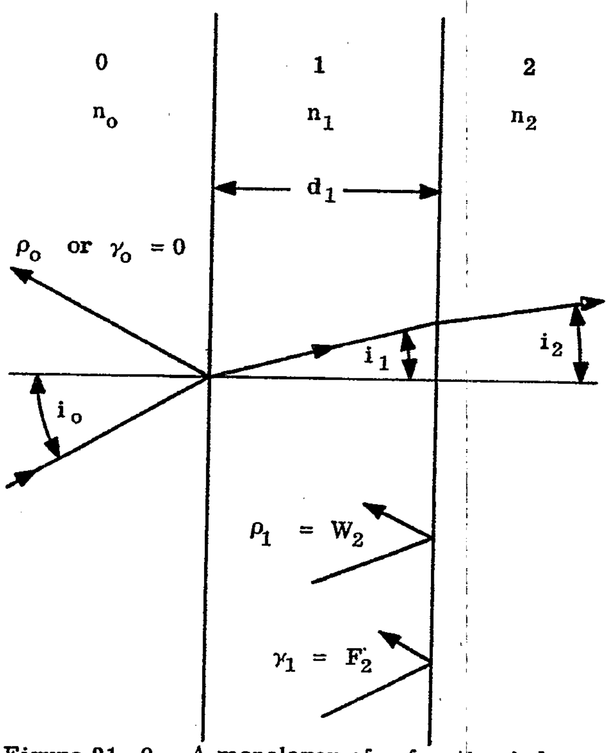


Figure 21.9- A monolayer of refractive index n_1 and thickness d_1 between two media having refractive indices n_o and n_2.

21.2.16.3 In dealing with the admittances Y_ν and y_ν, equations (79a) and (87a) show that the zero condition is

$$Y_1 = -M_o , \tag{99}$$

or

$$y_1 = m_o^2/M_o , \tag{99a}$$

according as the electric vector or the magnetic vector is perpendicular to the plane of incidence.

21.2.16.4 With respect to the Q-method, equation (93) shows that the zero condition is $Q_1 + (W_1^2 - 1) Q_2 = 0$, or

$$Q_1 = (1 - W_1^2) Q_2 , \tag{100}$$

when the electric vector is perpendicular to the plane of incidence.

21.3 ZERO REFLECTANCE FROM NON-ABSORBING MONOLAYERS AND SUBSTRATES

21.3.1 Introduction. The problem is to design a film that produces zero reflectance. The variables of the film are its refractive index, n_1, and its thickness, d_1. The usual restriction of the discussion to normal incidence will not be made because this restriction avoids too many pertinent and practical facts associated with oblique incidence. The following discussion will hinge upon the method of the complex reflectances. The complex reflectances ρ_o and γ_o are given by equations (96) and (97), respectively, with

$$\rho_1 = W_2 , \quad \text{and} \quad \gamma_1 = F_2 . \tag{101}$$

Equation (101) applies to monolayers, i.e. to cases $N = 1$. From equation (96), the energy reflectance, $|\rho_o|^2$, is given by

$$|\rho_o|^2 = \frac{W_1 \bar{W}_1 + W_2 \bar{W}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + W_1 \bar{W}_2 \overline{\exp(i\beta_1)} + \bar{W}_1 W_2 \exp(i\beta_1)}{1 + W_1 \bar{W}_1 W_2 \bar{W}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + W_1 W_2 \exp(i\beta_1) + \bar{W}_1 \bar{W}_2 \overline{\exp(i\beta_1)}} . \tag{102}$$

Similarly, from equations (101) and (97)

$$\gamma_o{}^2 = \frac{F_1 \bar{F}_1 + F_2 \bar{F}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + F_1 \bar{F}_2 \overline{\exp(i\beta_1)} + \bar{F}_1 F_2 \exp(i\beta_1)}{1 + F_1 \bar{F}_1 F_2 \bar{F}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + F_1 F_2 \exp(i\beta_1) + \bar{F}_1 \bar{F}_2 \overline{\exp(i\beta_1)}} . \tag{103}$$

21.3.2 Total internal reflection with M_2 pure imaginary. Let us consider first the class of cases in which $n_o p_o > n_2$ but in which n_1 is chosen so that $n_o p_o < n_1$. Then according to equation (19), M_1 is real but M_2 is pure imaginary. Consequently, from equations (20) and (21), W_1 and F_1 are real and W_2 and F_2 are complex such that

$$W_2 \bar{W}_2 = 1 ; \quad F_2 \bar{F}_2 = 1 . \tag{104}$$

Furthermore, $\beta_1 = 4\pi M_1 d_1/\lambda$ will be real. Equation (102) assumes now the simplified form

$$\rho_o{}^2 = \frac{1 + W_1 \bar{W}_1 + 2|W_1| |W_2| \cos\left[\beta_1 + \arg(W_2) - \arg(W_1)\right]}{1 + W_1 \bar{W}_1 + 2|W_1| |W_2| \cos\left[\beta_1 + \arg(W_2) + \arg(W_1)\right]} . \tag{105}$$

Because W_1 is real, $\arg(W_1) = 0$ and $|\rho_o|^2 = 1$. Similarly, $|\gamma_o|^2 = 1$ irrespective of the value of β_1. If n_1 is chosen so that $n_1^2 > n_o^2 p_o^2$, the film cannot alter the energy reflectance of the coated interface and the total reflection remains complete irrespective of the state of polarization of the incident beam. On the other hand, the phase change on reflection can be modified.

21.3.3 Total internal reflection with both M_1 and M_2 pure imaginary. Consider next the class of cases in which $n_o p_o$ exceeds both n_1 and n_2. Both M_1 and M_2 are then pure imaginary. Equations (20) and (21) now show that W_2 and F_2 are real but that W_1 and F_2 are complex such that

$$W_1 \bar{W}_1 = 1 ; \quad F_1 \bar{F}_1 = 1 . \tag{106}$$

Furthermore, since $\beta_1 = 4\pi M_1 d_1/\lambda$,

$$\exp(i\beta_1) = \overline{\exp(i\beta_1)} = \exp(-4\pi |n_o^2 p_o^2 - n_1^2|^{1/2} d_1/\lambda), \tag{107}$$

an attenuation factor that we shall designate temporarily by A. From equations (102), (106) and (107)

$$\rho_o{}^2 = \frac{1 + W_2^2 A^2 + 2A |W_1| W_2 \cos[\arg(W_1) - \arg(W_2)]}{1 + W_2^2 A^2 + 2A |W_1| W_2 \cos[\arg(W_1) + \arg(W_2)]} . \tag{108}$$

Since W_2 is real, $\arg(W_2) = 0, \pi, 2\pi$, etc. Hence $|\rho_o|^2 = 1$ irrespective of the thickness of the film. A similar conclusion holds when the magnetic vector is perpendicular to the plane of incidence. If i_o is chosen so that $n_o \sin i_o$ exceeds n_1 and n_2, the film cannot modify the energy reflectance of the coated interface.

21.3.4 Total internal reflectance when M_1 is pure imaginary. Total internal reflection may or may not occur when $n_o \sin i_o$ exceeds n_1 but not n_2. In this class of cases, M_1 is pure imaginary and M_2 is real. It can be seen from equations (20) and (21) that the Fresnel coefficients of reflection are complex such that

$$W_1 \overline{W}_1 = 1; \; W_2 \overline{W}_2 = 1; \; F_1 \overline{F}_1 = 1; \; F_2 \overline{F}_2 = 1. \tag{109}$$

In addition, β_1 obeys equation (107). Equation (102) assumes the form

$$|\rho_o|^2 = \frac{1 + A^2 + A(W_1 \overline{W}_2 + \overline{W}_1 W_2)}{1 + A^2 + A(W_1 W_2 + \overline{W}_1 \overline{W}_2)} , \tag{110}$$

in which $A = \exp(i\beta_1)$, a real attenuation factor. Because $A \to 0$ as $d_1 \to \infty$, $|\rho_o|^2 \to 1$ as the thickness d_1 of the film is increased. In fact, A falls very rapidly with increasing d_1. A relatively thin film can therefore produce almost total * internal reflection. One can show that, subject to equation (109),

$$W_1 \overline{W}_2 + \overline{W}_1 W_2 \leqq W_1 W_2 + \overline{W}_1 \overline{W}_2 . \tag{111}$$

Hence $|\rho_o|^2 \leqq 1$.

21.3.5 Zero reflectance; the E-vector perpendicular to the plane of incidence. Zero reflectance is possible with monolayers when $n_o \sin i_o$ is less than n_1 or n_2, i.e. when both M_1 and M_2 are real. Since $\rho_1 = W_2$ for monolayers (case $N = 1$), the zero condition of equation (98) becomes

$$W_2 e^{i\beta_1} = -W_1 . \tag{112}$$

Because W_1 and W_2 are real, equation (112) requires that $\exp(i\beta_1)$ be real. Two choices are possible

$$\beta_1 = \begin{cases} \mu\pi ; \; \mu \text{ an odd integer;} & (113) \\ \nu 2\pi ; \; \nu \text{ any integer;} & (113a) \end{cases}$$

$$\beta_1 = \frac{4\pi}{\lambda} n_1 d_1 \cos i_1 . \tag{113b}$$

We shall restrict our attention to the more interesting and important** choice (113). For all choices of μ, $\exp(i\beta_1) < 0$. Hence from equation (112) one must have

$$W_1 = W_2 . \tag{114}$$

Therefore from equation (20) and (114),

$$(M_o - M_1)(M_1 + M_2) = (M_o + M_1)(M_1 - M_2) ,$$

so that

$$M_1 = \sqrt{M_o M_2} . \tag{114a}$$

---

*Films belonging to the class discussed in this section are often said to be films that frustrate total internal reflection. For the simple case treated here, total reflection is frustrated more thoroughly as the thickness of the film is decreased.

**The choice (113a) leads to the theory of the "soap film." Thus at normal incidence the optical path $n_1 d_1$ is an integral number of half-wavelengths.

At normal incidence equations (113), (113b) and (114a) reduce to the pair of well known and independent conditions

$$n_1 d_1 = \mu \frac{\lambda}{4} , \tag{115}$$

where μ = odd, and $n_1 = \sqrt{n_o n_2}$.

At oblique incidence, it is often advantageous to assemble the independent conditions (113) and (114a) in the more explicit form

$$n_1 d_1 \cos i_1 = \mu \frac{\lambda}{4} \; ; \; \mu \text{ odd} ;$$

and (116)

$$\frac{n_1}{\sqrt{n_o n_2}} = \frac{\sqrt{\cos i_o \cos i_2}}{\cos i_1} ,$$

in which, from Snell's law, $n_o \sin i_o = n_1 \sin i_1 = n_2 \sin i_2$. The choice of n_1 and d_2 for zero reflectance depends therefore upon the angle, i_o, of incidence. The first condition of equation (115) is often called the quarter wave condition. The effective "interference path," $n_1 d_1 \cos i_1$, is equal to the optical path, $n_1 d_1$, of the film at normal incidence. The interference path, $n_1 d_1 \cos i_1$, always obeys the quarter wave condition.

21.3.6 Zero reflectance; the H-vector perpendicular to the plane of incidence. As in Section 21.3.5, zero reflectance is possible when both M_1 and M_2 are real; but, as we shall see, the zero condition corresponding to (114a) is significantly different. Since $\gamma_1 = F_2$ for monolayers, the zero condition (98a) assumes the form

$$F_2 \, e^{i\beta_1} = - F_1 . \tag{117}$$

With non-absorbing media, it will become clear from equation (21) that the Fresnel coefficients, F_1 and F_2, of reflection must be real when M_1 and M_2 are real. Conditions (113) and (113a) apply again. Corresponding to the choice (113), equation (117) requires that

$$F_1 = F_2 . \tag{118}$$

From equations (118) and (21) one obtains straightforwardly,

$$\frac{M_1}{\sqrt{M_o M_2}} = \frac{n_1^2}{n_o \, n_2} , \tag{118a}$$

a condition that should be compared with equation (114a). Upon introducing $M_1 = n_1 \cos i_1$, $M_2 = n_2 \cos i_2$ and $M_o = n_o \sin i_o$, one finds instead of equation (116) that

$$\frac{n_1}{\sqrt{n_o n_2}} = \frac{\cos i_1}{\sqrt{\cos i_o \cos i_2}} . \tag{119}$$

21.3.7 Summary.

21.3.7.1 We learn that a film that will produce zero reflectance at oblique incidence when the E-vector is perpendicular to the plane of incidence cannot be expected to produce zero reflectance when the H-vector is perpendicular to the plane of incidence. This conclusion could have been expected; for when the magnetic vector is perpendicular to the plane of incidence, the reflectance is automatically zero at Brewster's angle of incidence without the use of a film whereas,the reflectance is not zero when the electric vector is perpendicular to the plane of incidence. We conclude also that a monolayer cannot in principle produce strictly zero energy reflectance with unpolarized, incident light at other than normal incidence.

21.3.7.2 The reflectance method can be applied in a systematic manner to design bilayers, trilayers, etc. that produce zero reflectance at one or more wavelengths. Except with certain simplified and restricted combinations, the details of the analysis become exceedingly tedious as the number, N, of layers is increased beyond N = 3.

21.4 MATRIX METHODS

21.4.1 Introduction. Matrix methods possess distinct advantages for computation with desk or automatic calculators. For example, the nonlinear recursion formulae that relate successive interfacial reflectances are avoided. We shall construct matrix methods for computing the complex amplitudes, τ_ν and r_ν, for cases in which the electric vector is perpendicular to the plane of incidence, and for computing the complex amplitudes, T_ν and R_ν, for cases in which the magnetic vector is perpendicular to the plane of incidence. One may treat other states of polarization by splitting the field into two parts, in one of which the E-vector is perpendicular to the plane of incidence, and in the other of which the magnetic vector is perpendicular to the plane of incidence. The complex amplitudes, τ_ν and r_ν, retain the same physical significance as described in Sections 21.2.8 and 21.2.9. The complex amplitudes, T_ν and R_ν, retain the same significance as in Section 21.2.10.

21.4.2 Matrix methods; the E-vector perpendicular to the plane of incidence.

21.4.2.1 Equations (65) and (66) describe the electric vector propagated to the right and left, respectively, in the ν^{th} layer or medium, Figure 21.6. The electric vector has only the y-component. Thus,

$$\begin{aligned}(E_{y,\nu})_{\text{transmitted}} &= \tau_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu (z - L_\nu)} ,\\ (E_{y,\nu})_{\text{reflected}} &= r_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{-ikM_\nu (z - L_\nu)} .\end{aligned} \tag{120}$$

From equation (27) the corresponding tangential component $H_{x,\nu}$ of the magnetic vector is determined by

$$H_{x,\nu} = \frac{i}{k} \frac{\partial E_{y,\nu}}{\partial z} . \tag{121}$$

Hence,

$$\begin{aligned}(H_{x,\nu})_{\text{transmitted}} &= - M_\nu \; (E_{y,\nu})_{\text{transmitted}} ,\\ (H_{x,\nu})_{\text{reflected}} &= M_\nu \; (E_{y,\nu})_{\text{reflected}} .\end{aligned} \tag{122}$$

Let the total tangential components in the ν^{th} layer or medium be denoted by $H_{T,\nu}$ and $E_{T,\nu}$. Then, by definition,

$$\begin{aligned}E_{T,\nu} &= (E_{y,\nu})_{\text{transmitted}} + (E_{y,\nu})_{\text{reflected}} ,\\ H_{T,\nu} &= (H_{x,\nu})_{\text{transmitted}} + (H_{x,\nu})_{\text{reflected}} .\end{aligned} \tag{123}$$

The total tangential components $E_{T,\nu}$ and $H_{T,\nu}$ are continuous across every interface of the multilayer. From equations (120) and (122), $E_{T,\nu}$ and $H_{T,\nu}$ are continuous across the interface $z = L_{\nu-1}$ provided that

$$r_{\nu-1} + \tau_{\nu-1} = r_\nu \; e^{i\frac{\beta_\nu}{2}} + \tau_\nu \; e^{-i\frac{\beta_\nu}{2}} , \tag{124}$$

$$r_{\nu-1} - \tau_{\nu-1} = \frac{M_\nu}{M_{\nu-1}} \left[r_\nu \; e^{i\frac{\beta_\nu}{2}} - \tau_\nu \; e^{-i\frac{\beta_\nu}{2}} \right] , \tag{124a}$$

since $k\, M_\nu\, (L_\nu - L_{\nu-1}) = k\, M_\nu\, d_\nu = \beta_\nu/2$. Therefore,

$$2\, r_{\nu-1} = \frac{r_\nu}{M_{\nu-1}} \; e^{i\frac{\beta_\nu}{2}} (M_{\nu-1} + M_\nu) + \frac{\tau_\nu}{M_{\nu-1}} \; e^{-i\frac{\beta_\nu}{2}} (M_{\nu-1} - M_\nu) , \tag{125}$$

$$2\, \tau_{\nu-1} = \frac{r_\nu}{M_{\nu-1}} \; e^{i\frac{\beta_\nu}{2}} (M_{\nu-1} - M_\nu) + \frac{\tau_\nu}{M_{\nu-1}} \; e^{-i\frac{\beta_\nu}{2}} (M_{\nu-1} + M_\nu) . \tag{125a}$$

It may be noted that division of equations (125) leads to the result of equation (57).

21.4.2.2 The following matrix algebra is all that is needed in executing the matrix method. A matrix describes a linear transformation from one pair of variables x_1, y_1 to a second pair x_2, y_2. Thus the linear transformation

$$\begin{aligned}x_2 &= a_{11}\, x_1 + a_{12}\, y_1 ,\\ y_2 &= a_{21}\, x_1 + a_{22}\, y_1 ,\end{aligned} \tag{126}$$

is written in matrix notation as

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} . \tag{127}$$

Suppose that a further transformation to the variables x_3 , y_3 is given by

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} . \tag{128}$$

One can verify by eliminating x_2 , y_2 that the matrix product of equation (128) is a matrix such that

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} b_{11}\, a_{11} + b_{12}\, a_{21} & b_{11}\, a_{12} + b_{12}\, a_{22} \\ b_{21}\, a_{11} + b_{22}\, a_{21} & b_{21}\, a_{12} + b_{22}\, a_{22} \end{bmatrix} . \tag{129}$$

Equation (129) gives the rule for multiplication. In performing the multiplication $[b_{kl}] \times [a_{mn}]$, to obtain the element in row i and column j of the product matrix take the scalar product of the i^{th} row of matrix $\underline{b}$, and the j^{th} column of matrix $\underline{a}$. The continued product of any number of matrices can be performed by repeating the rule. Multiplication is not commutative, i.e. $[b_{kl}] \times [a_{mn}] \neq [a_{mn}] \times [b_{kl}]$.

21.4.2.3 Returning to equations (125), we observe that in matrix notation

$$\begin{bmatrix} r_{\nu-1} \\ \tau_{\nu-1} \end{bmatrix} = \frac{\mathcal{M}_\nu}{2\, M_{\nu-1}} \begin{bmatrix} r_\nu \\ \tau_\nu \end{bmatrix} \tag{130}$$

where $\mathcal{M}_\nu$ denotes the square matrix

$$\mathcal{M}_\nu = \begin{bmatrix} (M_{\nu-1} + M_\nu)\, e^{i \frac{\beta\nu}{2}} & (M_{\nu-1} - M_\nu)\, e^{-i \frac{\beta\nu}{2}} \\ (M_{\nu-1} - M_\nu)\, e^{i \frac{\beta\nu}{2}} & (M_{\nu-1} + M_\nu)\, e^{-i \frac{\beta\nu}{2}} \end{bmatrix} . \tag{130a}$$

Therefore

$$\begin{bmatrix} r_{\nu-1} \\ \tau_{\nu-1} \end{bmatrix} = \frac{\mathcal{M}_\nu}{2\, M_{\nu-1}} \; \frac{\mathcal{M}_{\nu+1}}{2\, M_\nu} \begin{bmatrix} r_{\nu+1} \\ \tau_{\nu+1} \end{bmatrix} , \tag{131}$$

whence,

$$\begin{bmatrix} r_{\nu-1} \\ \tau_{\nu-1} \end{bmatrix} = \frac{1}{2^{N-\nu+1}} \prod_{j=\nu-1}^{N-1} \frac{1}{M_j} \prod_{j=\nu}^{N-1} \mathcal{M}_j \begin{bmatrix} r_N \\ \tau_N \end{bmatrix} , \tag{131a}$$

and

$$\begin{bmatrix} r_o \\ \tau_o \end{bmatrix} = \frac{1}{2^N} \prod_{j=0}^{N-1} \frac{1}{M_j} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} r_N \\ \tau_N \end{bmatrix} . \tag{131b}$$

But r_N / τ_N is the Fresnel coefficient of reflection at the last interface, so that

$$r_N = \tau_N\, W_{N+1} = \tau_N \frac{M_N - M_{N+1}}{M_N + M_{N+1}} . \tag{132}$$

Furthermore, τ_{N+1}/τ_N is the Fresnel coefficient of transmission of the last interface. Hence,

$$\tau_N = \frac{M_N + M_{N+1}}{2\, M_N} \tau_{N+1} ; \tag{132a}$$

$$r_N = \frac{M_N - M_{N+1}}{2\, M_N} \tau_{N+1} ; \tag{132b}$$

$$\begin{bmatrix} r_N \\ \tau_N \end{bmatrix} = \frac{1}{2\, M_N} \begin{bmatrix} (M_N - M_{N+1})\; \tau_{N+1} \\ (M_N + M_{N+1})\; \tau_{N+1} \end{bmatrix} = \frac{\tau_{N+1}}{2\, M_N} \begin{bmatrix} (M_N - M_{N+1}) \\ (M_N + M_{N+1}) \end{bmatrix} . \tag{132c}$$

Finally, from equations (131b) and (132c), one obtains

$$\begin{bmatrix} r_o \\ \tau_o \end{bmatrix} = \frac{\tau_{N+1}}{2^{N+1}} \prod_{j=0}^{N} \frac{1}{M_j} \prod_{j=1}^{N} \mathfrak{M}_j \begin{bmatrix} (M_N - M_{N+1}) \\ (M_N + M_{N+1}) \end{bmatrix} . \tag{133}$$

In this equation the unknowns are usually r_o and τ_{N+1}. τ_o is a complex number that specifies the amplitude and phase of the incident electric vector at $x = 0$ at the left hand side of the first interface $z = 0$, Figure 21.6. One may set $\tau_o = 1$. r_o specifies the amplitude and phase of the reflected vector at $x = 0$ at the left hand side of the first interface of the multilayer. τ_{N+1} specifies the amplitude and phase of the transmitted electric vector at $x = 0$ at the right hand side of the last interface $z = L_N$, Figure 21.6. N is the number of layers.

21.4.2.4 One important advantage of this matrix method is that it enables the exploration of the effects of changing the thickness and refractive index of the ν^{th} layer without recomputing the entire matrix product beyond the $(\nu-1)^{th}$ layer. Changing thickness of the νth layer alters only the matrix $\mathfrak{M}_\nu$. Because

$$\prod_{j=1}^{N} \mathfrak{M}_j = \prod_{j=1}^{\nu-1} \mathfrak{M}_j \times \mathfrak{M}_\nu \times \prod_{j=\nu+1}^{N} \mathfrak{M}_j , \tag{134}$$

the first and third matrix products in the right hand member can be computed as matrices that remain fixed during the exploration of the effect of changing thickness. Changing refractive index of the ν^{th} layer alters both $\mathfrak{M}_\nu$ and $\mathfrak{M}_{\nu+1}$. In this case one utilizes instead of equation (134),

$$\prod_{j=1}^{N} \mathfrak{M}_j = \prod_{j=1}^{\nu-1} \mathfrak{M}_j \times \mathfrak{M}_\nu \mathfrak{M}_{\nu+1} \times \prod_{j=\nu+2}^{N} \mathfrak{M}_j . \tag{135}$$

This case illustrates also the manipulation of sub-products of matrices where these sub-products may remain fixed or may be varied. A particular fixed sub-product may be shifted to different positions in the complete matrix product -- corresponding to shifting a group of layers in the multilayer. The use of the matrix method for designing multilayers with periodic structure has been discussed by W. Weinstein.(5)

21.4.2.5 A further interpretation of the matrices $\mathfrak{M}_\nu$ is appropriate. It is not a trivial fact that the matrix $\mathfrak{M}_\nu$ of equation (130a) can be expressed as the matrix product

$$\mathfrak{M}_\nu = \begin{bmatrix} M_{\nu-1} + M_\nu & M_{\nu-1} - M_\nu \\ M_{\nu-1} - M_\nu & M_{\nu-1} + M_\nu \end{bmatrix} \begin{bmatrix} e^{i\frac{\beta\nu}{2}} & 0 \\ 0 & e^{-i\frac{\beta\nu}{2}} \end{bmatrix} , \tag{136}$$

$$= \mathcal{S}_\nu \mathcal{T}_\nu \tag{136a}$$

wherein $\mathcal{S}_\nu$ is the first right hand matrix in equation (136) and $\mathcal{T}_\nu$ is the second right hand matrix. $\mathcal{S}_\nu$ and $\mathcal{T}_\nu$ are matrices that depend, respectively, upon the optical constants, M_ν, and the interference path, β_ν, of the νth layer. With reference to equation (133), one may write when desired

$$\prod_{j=1}^{N} \mathfrak{M}_j = \prod_{j=1}^{N} \mathcal{S}_j \mathcal{T}_j \tag{136b}$$

21.4.2.6 Let us now consider solution in terms of the amplitude-sums. With respect to equations (124), let

$$\begin{aligned} A_\nu &= r_\nu + \tau_\nu ; \\ B_\nu &= M_\nu (r_\nu - \tau_\nu) . \end{aligned} \tag{137}$$

Then

$$\begin{aligned} r_\nu &= \frac{1}{2}\left(A_\nu + \frac{B_\nu}{M_\nu} \right) ; \\ \tau_\nu &= \frac{1}{2}\left(A_\nu - \frac{B_\nu}{M_\nu} \right) . \end{aligned} \tag{137a}$$

By eliminating r_ν and τ_ν from equations (124) and (124a) with the aid of equation (137a), one finds that

$$\begin{aligned} A_{\nu-1} &= A_\nu \cos(\beta_\nu/2) + B_\nu \frac{i \sin(\beta_\nu/2)}{M_\nu} ; \\ B_{\nu-1} &= A_\nu \, i M_\nu \sin(\beta_\nu/2) + B_\nu \cos(\beta_\nu/2) . \end{aligned} \tag{137b}$$

---

(5) Walter Weinstein, Vacuum, 4, 3-18 (1954).

Hence the amplitude-sums, A_ν and B_ν, obey the relation

$$\begin{bmatrix} A_{\nu-1} \\ B_{\nu-1} \end{bmatrix} = \begin{bmatrix} \cos(\beta_\nu/2) & \dfrac{i \sin(\beta_\nu/2)}{M_\nu} \\ i\, M_\nu \sin(\beta_\nu/2) & \cos(\beta_\nu/2) \end{bmatrix} \begin{bmatrix} A_\nu \\ B_\nu \end{bmatrix} . \tag{137c}$$

From equation (132)

$$\begin{aligned} A_N &= r_N + \tau_N = \tau_{N+1}; \\ B_N &= M_N (r_N - \tau_N) = - M_{N+1}\, \tau_{N+1}. \end{aligned} \tag{137d}$$

Therefore

$$\begin{bmatrix} A_o \\ B_o \end{bmatrix} = \tau_{N+1} \prod_{\nu=1}^{N} \begin{bmatrix} \cos(\beta_\nu/2) & \dfrac{i \sin(\beta_\nu/2)}{M_\nu} \\ i\, M_\nu \sin(\beta_\nu/2) & \cos(\beta_\nu/2) \end{bmatrix} \begin{bmatrix} 1 \\ -M_{N+1} \end{bmatrix} . \tag{137e}$$

The method of computation based upon the use of equations (137e) and (137a) is especially advantageous in dealing with non-absorbing multilayers and substrates; for then all M_ν and β_ν are real. Computation of the matrix product of equation (137e) becomes relatively simple. It should be noted that the determinant of each matrix is unity. As pointed out by W. Weinstein,(6) the determinant of the product of matrices is the product of their determinants. Therefore the determinant of products of these matrices is unity -- a valuable fact for the purpose of checking calculations. The complex amplitudes, r_o and τ_{N+1}, are computed at the end, with τ_o assigned a convenient value such as unity. The reflectance, $\rho_o = r_o/\tau_o$, and the transmittance, τ_{N+1}/τ_o, become known.

21.4.3 Matrix methods; the H-vector perpendicular to the plane of incidence.

21.4.3.1 When the magnetic vector is perpendicular to the plane of incidence, the tangential components of E and H are given by equations (84). As in equation (123), we form the total tangential components consisting of the transmitted and reflected waves. Application of the continuity condition for the total tangential components of E and H at the interface $z = L_{\nu-1}$ leads, as in equation (124), to the result

$$R_{\nu-1} + T_{\nu-1} = R_\nu\, e^{i\frac{\beta\nu}{2}} + T_\nu\, e^{-i\frac{\beta\nu}{2}}; \tag{138}$$

$$R_{\nu-1} - T_{\nu-1} = \frac{M_\nu\, m_{\nu-1}^2}{M_{\nu-1}\, m_\nu^2} \left[R_\nu\, e^{i\frac{\beta\nu}{2}} - T_\nu\, e^{-i\frac{\beta\nu}{2}} \right]. \tag{138a}$$

Therefore

$$2 R_{\nu-1} = \frac{R_\nu\, e^{i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} (M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2) + \frac{T_\nu\, e^{-i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} (M_{\nu-1}\, m_\nu^2 - M_\nu\, m_{\nu-1}^2) \tag{139}$$

$$2 T_{\nu-1} = \frac{R_\nu\, e^{i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} (M_{\nu-1}\, m_\nu^2 - M_\nu\, m_{\nu-1}^2) + \frac{T_\nu\, e^{-i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} (M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2). \tag{139a}$$

Hence,

$$\begin{bmatrix} R_{\nu-1} \\ T_{\nu-1} \end{bmatrix} = \frac{\mathcal{M}_\nu}{2\, M_{\nu-1}\, m_\nu^2} \begin{bmatrix} R_\nu \\ T_\nu \end{bmatrix}, \tag{140}$$

in which $\mathcal{M}_\nu$ is the matrix

$$\mathcal{M}_\nu = \begin{bmatrix} (M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2)\, e^{i\frac{\beta\nu}{2}} & (M_{\nu-1}\, m_\nu^2 - M_\nu\, m_{\nu-1}^2)\, e^{-i\frac{\beta\nu}{2}} \\ (M_{\nu-1}\, m_\nu^2 - M_\nu\, m_{\nu-1}^2)\, e^{i\frac{\beta\nu}{2}} & (M_{\nu-1}\, m_\nu^2 + M_\nu\, m_{\nu-1}^2)\, e^{-i\frac{\beta\nu}{2}} \end{bmatrix}. \tag{140a}$$

---

(6) ibid p 8

Therefore

$$\begin{bmatrix} R_o \\ T_o \end{bmatrix} = \frac{1}{2^N} \prod_{j=0}^{N-1} \frac{1}{M_j \, m_{j+1}^2} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} R_N \\ T_N \end{bmatrix} , \qquad (140b)$$

however,

$$T_N = \frac{m_{N+1}^2 \, M_N + m_N^2 \, M_{N+1}}{2 \, M_N \, m_{N+1}^2} \; T_{N+1} , \qquad (140c)$$

and

$$R_N = F_N \; T_N = \frac{m_{N+1}^2 \, M_N - m_N^2 \, M_{N+1}}{2 \, M_N \, m_{N+1}^2} \; T_{N+1} , \qquad (140d)$$

with F_N given by equation (21). Hence,

$$\begin{bmatrix} R_o \\ T_o \end{bmatrix} = \frac{T_{N+1}}{2^{N+1}} \prod_{j=0}^{N} \frac{1}{M_j \, m_{j+1}^2} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} m_{N+1}^2 \, M_N - m_N^2 \, M_{N+1} \\ m_{N+1}^2 \, M_N + m_N^2 \, M_{N+1} \end{bmatrix} , \qquad (141)$$

in which the matrices $\mathcal{M}_j$ are given by equation (140a). The complex amplitudes R_ν and T_ν retain the same physical significance as in 21.2.10.

21.4.3.2 Equation (141) serves to determine T_{N+1} from T_o and R_o from T_o and T_{N+1}. In most problems, one may set $T_o = 1$. Finally, the complex reflectance γ_o is computed from its definition $\gamma_o = R_o/T_o$. It will be observed that equations (133) and (141) determine the complex transmittances, τ_{N+1} and T_{N+1}, of the multilayer quite directly without necessitating additional multiplications such as those of equation (54) or (70).

21.4.3.3 The solution in terms of the amplitude sums can be obtained as follows. Let

$$\begin{aligned} C_\nu &= R_\nu + T_\nu ; \\ D_\nu &= \frac{M_\nu}{m_\nu^2} \, (R_\nu - T_\nu) . \end{aligned} \qquad (142)$$

Then

$$\begin{aligned} R_\nu &= \frac{1}{2} \left(C_\nu + \frac{m_\nu^2}{M_\nu} \, D_\nu \right) ; \\ T_\nu &= \frac{1}{2} \left(C_\nu - \frac{m_\nu^2}{M_\nu} \, D_\nu \right) . \end{aligned} \qquad (142a)$$

By eliminating R_ν and T_ν from equation (138) with the aid of equation (142a), one obtains

$$\begin{aligned} C_{\nu-1} &= C_\nu \cos \frac{\beta_\nu}{2} + i \, D_\nu \frac{m_\nu^2}{M_\nu} \sin \frac{\beta_\nu}{2} ; \\ D_{\nu-1} &= i \, C_\nu \frac{M_\nu}{m_\nu^2} \sin \frac{\beta_\nu}{2} + D_\nu \cos \frac{\beta_\nu}{2} . \end{aligned} \qquad (142b)$$

Therefore

$$\begin{bmatrix} C_{\nu-1} \\ D_{\nu-1} \end{bmatrix} = \begin{bmatrix} \cos \dfrac{\beta_\nu}{2} & i \, \dfrac{m_\nu^2}{M_\nu} \sin \dfrac{\beta_\nu}{2} \\ i \, \dfrac{M_\nu}{m_\nu^2} \sin \dfrac{\beta_\nu}{2} & \cos \dfrac{\beta_\nu}{2} \end{bmatrix} \begin{bmatrix} C_\nu \\ D_\nu \end{bmatrix} . \qquad (142c)$$

From equations (140c) and (140d),

$$C_N = R_N + T_N = T_{N+1};$$
$$D_N = \frac{M_N}{m_N^2}(R_N - T_N) = -\frac{M_{N+1}}{m_{N+1}^2} T_{N+1}. \qquad (142d)$$

Therefore, upon forming $\begin{bmatrix} C_o \\ D_o \end{bmatrix}$ from equation (142c), one obtains

$$\begin{bmatrix} C_o \\ D_o \end{bmatrix} = T_{N+1} \prod_{\nu=1}^{N} \begin{bmatrix} \cos\frac{\beta\nu}{2} & i\,\frac{m_\nu^2}{M_\nu}\sin\frac{\beta\nu}{2} \\ i\,\frac{M\nu}{m_\nu^2}\sin\frac{\beta\nu}{2} & \cos\frac{\beta\nu}{2} \end{bmatrix} \begin{bmatrix} 1 \\ -M_{N+1}/m_{N+1}^2 \end{bmatrix}, \qquad (142e)$$

a result that should be compared with equation (137e). The remarks in the paragraph following equation (137e) apply again to equation (142e).

21. 5 QUATERNION METHODS

21. 5. 1 Introduction. A computing program based upon quarternions instead of matrices has been introduced by Dr. Gordon L. Walker * for analyzing thin films with the aid of automatic computers. The relative advantages of matrices and quaternions depend mainly upon circumstances at the location of the automatic calculator. For example, wherever a programmed matrix formulation is available, it may be simpler to adapt a matrix method. As has been seen in Section 21. 4, it is both natural and direct to state the solution to the problems of thin films in matrix notation. Because many solutions in matrix form have been given, we shall restrict our considerations to showing how any matrix solution can be transformed into the corresponding quaternion form.

10. 5. 2 Quaternions. A quaternion Q is the sum of a scalar and a vector. Thus,

$$Q = q_o\sigma_o + q_1\sigma_1 + q_2\sigma_2 + q_3\sigma_3, \qquad (143)$$

in which σ_1, σ_2 and σ_3 are unit vectors and $\sigma_o = 1$. Coefficients q_o, q_1, q_2 and q_3 are scalars that can be complex imaginary. With respect to summation,

$$P + Q = Q + P = (p_o + q_o)\sigma_o + (p_1 + q_1)\sigma_1 + (p_2 + q_2)\sigma_2 + (p_3 + q_3)\sigma_3. \qquad (143a)$$

If b is a scalar,

$$bQ = Qb = bq_o\sigma_o + bq_1\sigma_1 + bq_2\sigma_2 + bq_3\sigma_3. \qquad (143b)$$

In forming the product of two quaternions, P and Q, one observes the following rules of operation with respect to the unit vectors.

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = -1; \qquad \sigma_o^2 = 1;$$
$$\sigma_1\sigma_2 = -\sigma_2\sigma_1 = \sigma_3;$$
$$\sigma_2\sigma_3 = -\sigma_3\sigma_2 = \sigma_1;$$
$$\sigma_3\sigma_1 = -\sigma_1\sigma_3 = \sigma_2. \qquad (143c)$$

It follows that $PQ \neq QP$. Let p_ν and q_ν be the coefficients of quaternions P and Q, respectively. Set

$$PQ = R = R_o\sigma_o + R_1\sigma_1 + R_2\sigma_2 + R_3\sigma_3. \qquad (143d)$$

---

*This unpublished scheme has been used by Drs. G. L. Walker, H. Jupnik and A. Traub at the American Optical Company. A method that resembles the method of quaternions in several respects has been discussed by M. Andre Herpin, Comptes Rendus Acad. Sci., 225, 182-183 (1947).

Then,

$$R_o = p_o q_o - p_1 q_1 - p_2 q_2 - p_3 q_3 ;$$
$$R_1 = p_o q_1 + q_o p_1 + p_2 q_3 - p_3 q_2 ;$$
$$R_2 = p_o q_2 + q_o p_2 + p_3 q_1 - p_1 q_3 ;$$
$$R_3 = p_o q_3 + q_o p_3 + p_1 q_2 - p_2 q_1 . \qquad (143e)$$

21.5.3 Corresponding matrices and quaternions. Let

$$\mathcal{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \qquad (144)$$

be a given matrix. Let $\tilde{\mathcal{M}}$ denote the corresponding quaternion. Then

$$\tilde{\mathcal{M}} = m_o \sigma_o + m_1 \sigma_1 + m_2 \sigma_2 + m_3 \sigma_3 , \qquad (144a)$$

wherein

$$m_o = \frac{m_{11} + m_{22}}{2} ; \qquad m_2 = \frac{m_{12} - m_{21}}{2} ;$$
$$m_1 = i \frac{m_{11} - m_{22}}{2} ; \qquad m_3 = i \frac{m_{12} + m_{21}}{2} . \qquad (144b)$$

On the other hand, let $\tilde{\mathcal{M}}$ be the given quaternion. In order to obtain the corresponding matrix, one replaces $\sigma_o = 1$ and the unit vectors by the matrices

$$\sigma_o = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ; \qquad \sigma_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} ;$$
$$\sigma_1 = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} ; \qquad \sigma_3 = \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} ; \qquad (144c)$$

in which σ_o is a unit matrix. Thus with $\tilde{\mathcal{M}}$ regarded as the given quaternion, the corresponding matrix is

$$\mathcal{M} = m_o \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + m_1 \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} + m_2 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + m_3 \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} ;$$

$$= \begin{bmatrix} m_o - i m_1 & m_2 - i m_3 \\ -(m_2 + i m_3) & m_o + i m_1 \end{bmatrix} . \qquad (144d)$$

The matrices of equations (144c) satisfy the requirements of equations (143c). The quaternion corresponding to the product of two matrices taken in a specified order is the product of the quaternions corresponding to each of the two matrices taken in the same order. Thus

$$\text{Quaternion } (\mathcal{M}_1 \mathcal{M}_2) = \tilde{\mathcal{M}}_1 \tilde{\mathcal{M}}_2 . \qquad (144e)$$

Similarly,

$$\text{Matrix } (\tilde{\mathcal{M}}_1 \tilde{\mathcal{M}}_2) = \mathcal{M}_1 \mathcal{M}_2 . \qquad (144f)$$

For the purposes of this text, equation (144e) is far more important than equation (144f). Repeated application of equation (144e) shows that

$$\text{Quaternion } \left(\prod_{\nu=1}^{N} \mathcal{M}_\nu \right) = \prod_{\nu=1}^{N} \tilde{\mathcal{M}}_\nu , \qquad (144g)$$

in which $\tilde{\mathcal{M}}_\nu$ is the quaternion that corresponds to matrix $\mathcal{M}_\nu$.

21.5.4 Replacements for matrix equations. Let $\mathcal{M}$ be the square matrix

$$\mathcal{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \tag{145}$$

that connects the quantities r, τ, A and B according to the law

$$\begin{bmatrix} r \\ \tau \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}, \tag{145a}$$

as in equations (126) and (127). One can verify almost directly from equations (126) to (129) that it is permissible to set

$$\begin{bmatrix} r \\ \tau \end{bmatrix} = \begin{bmatrix} r & 0 \\ \tau & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix}.$$

Hence equation (145a) can be written in the form

$$\begin{bmatrix} r & 0 \\ \tau & 0 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix}. \tag{145b}$$

The quaternion form of matrix equation (145b) is now obtained by replacing each of the three square matrices by its corresponding quaternion with the aid of equation (144b). Let U denote the quaternion corresponding to the left hand member of equation (145b), i.e. let

$$U = \text{Quaternion} \begin{bmatrix} r & 0 \\ \tau & 0 \end{bmatrix}. \tag{145c}$$

By applying the correspondence rules of equation (144b) to equation (145c), one finds that

$$U = \frac{r}{2}\sigma_o + i\frac{r}{2}\sigma_1 - \frac{\tau}{2}\sigma_2 + i\frac{\tau}{2}\sigma_3. \tag{145d}$$

Similarly, with respect to

$$S_o = \text{Quaternion} \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix},$$

$$S_o = \frac{A}{2}\sigma_o + i\frac{A}{2}\sigma_1 - \frac{B}{2}\sigma_2 + i\frac{B}{2}\sigma_3. \tag{145e}$$

Matrix equation (145b) is therefore replaced by the quaternion equation

$$U = \tilde{\mathcal{M}} S_o, \tag{145f}$$

in which

$$\tilde{\mathcal{M}} = \text{Quaternion} \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

$$= m_o\sigma_o + m_1\sigma_1 + m_2\sigma_2 + m_3\sigma_3, \tag{145g}$$

as in equations (144a) and (144b). It is instructive to examine the quaternion product $\tilde{\mathcal{M}}$ S_o. One finds from equations (143d), (143e), (145e) and (145g) that with

$$U = \tilde{\mathcal{M}} S_o = \sigma_o U_o + U_1\sigma_1 + U_2\sigma_2 + U_3\sigma_3, \tag{145h}$$

$$\begin{aligned} U_o &= \frac{A}{2}(m_o - i\,m_1) + \frac{B}{2}(m_2 - i\,m_3); \\ U_2 &= \frac{A}{2}(m_2 + i\,m_3) - \frac{B}{2}(m_o + i\,m_1); \end{aligned} \tag{145i}$$

together with

$$U_1 = i\,U_o; \quad U_3 = -i\,U_2. \tag{145j}$$

The relations between U_o, U_2 and m_{kl} are obtained from equations (144b) and (145i). One finds that

$$U_o = \frac{A}{2} m_{11} + \frac{B}{2} m_{12} ; \qquad U_2 = \frac{A}{2} m_{21} - \frac{B}{2} m_{22} . \tag{145k}$$

21.5.5 Quaternion solutions for r and τ. The solutions for r and τ are found by comparing equations (145d) and (145h). These equations require that

$$r = 2 U_o ; \qquad \tau = -2 U_2 ; \tag{146}$$

in which U_o and U_2 are the coefficients of σ_o and σ_2 in the quaternion product $\tilde{\mathcal{M}} S_o$ of equation (145f). Equation (146) forms a simple way of computing r and τ once U_o and U_2 have been found.

21.5.6 A check on the quaternion method. Equations (146) and (145k) can be combined to form a simple check on the correctness of the quaternion method. One finds directly that

$$r = m_{11} A + m_{12} B ; \qquad \tau = m_{21} A + m_{22} B . \tag{147}$$

These solutions for r and τ are those of the matrix equation (145a). If therefore equation (145a) is correct, equation (146) is correct.

21.5.7 A more useful statement of the formulation. In stating the matrix that corresponds to a multilayer, it is rarely convenient to specify the matrix in the form of equation (145a). Instead, it is convenient to express the matrix in the form

$$\begin{bmatrix} r \\ \tau \end{bmatrix} = F \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} \tag{148}$$

in which factor F is a scalar. For example, in the matrix of equation (137e) the factor $F = \tau_{N+1}$. Re-examination of the argument leading from equation (145a) to (146) shows that one obtains instead of equation (146) the slightly modified result

$$r = 2 F U_o ; \qquad \tau = -2 F U_2 ; \tag{149}$$

in which U_o and U_2 are the coefficients of σ_o and σ_2, respectively, of the quaternion corresponding to the matrix product of equation (148).

21.5.8 Special conditions. Depending upon the nature of the multilayer and upon the manner in which the corresponding matrix problem has been solved, special conditions may exist among the matrix elements m of the matrix $\mathcal{M}$ associated with the multilayer. For example, with respect to the matrix

$$\mathcal{M} = \prod_{\nu=1}^{N} \mathcal{M}_\nu = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} , \tag{150}$$

of equation (133), it follows that at normal incidence upon a system of non-absorbing layers

$$m_{22} = \bar{m}_{11} ; \qquad m_{21} = \bar{m}_{12} ; \tag{151}$$

because the diagonal elements of $\mathcal{M}_\nu$ of equation (130a) are then conjugate complex numbers. Correspondingly, from equation (144b)

$$m_o = R_e (m_{11}) ; \qquad m_2 = i \mathcal{I}_m (m_{12}) ; \qquad m_1 = -\mathcal{I}_m (m_{11}) ; \qquad m_3 = i R_e (m_{12}) . \tag{152}$$

It will not, however, be a purpose of this discussion to enumerate the various special conditions together with their consequences.

21.5.9 Application to the matrix solution of equation (133). The matrix solution of equation (133) applies to cases in which the electric vector is polarized so as to vibrate at right angles to the plane of incidence. This example will illustrate procedures that may be followed in converting any matrix solution for thin films into a solution in terms of quaternions. With respect to efficiency, it is undoubtedly preferable to choose as factor F the product

$$\frac{\tau_{N+1}}{2^{N+1}} \prod_{\nu=0}^{N} \frac{1}{M_\nu}$$

of equation (133). Let us demand, however, that the Fresnel coefficients W_ν of equation (20) shall appear as parameters. This can be accomplished by writing $\mathcal{M}_\nu$ of equation (130a) in the form

$$\mathcal{M}_\nu = (M_{\nu-1} + M_\nu) \begin{bmatrix} e^{i\frac{\beta\nu}{2}} & W_\nu\, e^{-i\frac{\beta\nu}{2}} \\ W_\nu\, e^{i\frac{\beta\nu}{2}} & e^{-i\frac{\beta\nu}{2}} \end{bmatrix} . \tag{153}$$

Let also the factor $M_N + M_{N+1}$ be removed from the second right hand matrix of equation (133). Then,

$$\begin{bmatrix} r_o \\ \tau_o \end{bmatrix} = F \prod_{\nu=1}^{N} \begin{bmatrix} e^{i\frac{\beta\nu}{2}} & W_\nu\, e^{-i\frac{\beta\nu}{2}} \\ W_\nu\, e^{i\frac{\beta\nu}{2}} & e^{-i\frac{\beta\nu}{2}} \end{bmatrix} \begin{bmatrix} W_{N+1} \\ 1 \end{bmatrix} ; \tag{154}$$

in which

$$F = \frac{\tau_{N+1}}{2^{N+1}} \; \frac{M_N + M_{N+1}}{M_o} \prod_{\nu=1}^{N} \frac{M_\nu + M_{\nu-1}}{M_\nu} . \tag{155}$$

We take $\tilde{\mathcal{M}}_\nu$ as the quaternion corresponding to the ν^{th} matrix of the product from $\nu = 1$ to $\nu = N$ of equation (154). Explicitly, we take

$$\tilde{\mathcal{M}}_\nu = \cos\frac{\beta\nu}{2}\,\sigma_o - \sin\frac{\beta\nu}{2}\,\sigma_1 - i\,W_\nu \sin\frac{\beta\nu}{2}\,\sigma_2 + i\,W_\nu \cos\frac{\beta\nu}{2}\,\sigma_3 . \tag{156}$$

The quaternion S_o of the last matrix of equation (154) is obtained from equation (145e) by setting

$$A = W_{N+1}; \qquad B = 1 . \tag{157}$$

We form and compute the quaternion

$$\tilde{\mathcal{M}} = \prod_{\nu=1}^{N} \tilde{\mathcal{M}}_\nu \tag{158}$$

with $\tilde{\mathcal{M}}_\nu$ given in terms of the physical properties β_ν and W_ν of the ν^{th} layer by equation (156). Next, as in equation (145f), we compute the quaternion U as the product

$$U = \frac{1}{2}\,\tilde{\mathcal{M}} \left[W_{N+1}\,\sigma_o + i\,W_{N+1}\,\sigma_1 - \sigma_2 + i\,\sigma_3 \right] ; \tag{159}$$

$$= \prod_{\nu=1}^{N} \tilde{\mathcal{M}}_\nu \; S_o$$

In making the last computation of equation (159), it is necessary to compute only U_o and U_2, the coefficients of σ_o and σ_2, respectively. With U_o and U_2 thus determined, it follows from equation (149) that

$$\begin{aligned} r_o &= 2\,F\,U_o ; \\ \tau_o &= -2\,F\,U_2 ; \end{aligned} \tag{160}$$

in which F is given by equation (155).

21.5.10 Determination of the complex reflectance and transmittance of the multilayer. The complex reflectance of the multilayer is given by the ratio $\rho_o = r_o / \tau_o$. From equation (160)

$$\rho_o = -\,U_o / U_2 , \tag{161}$$

a result that is independent of the factor F and that is evaluated at the right hand boundary of the medium of incidence with the electric vector perpendicular to the plane of incidence. The complex transmittance of the multilayer is given by the ratio τ_{N+1}/τ_0. From equations (160) and (155)

$$\tau_{N+1}/\tau_0 = -2^{N+1} \frac{M_0}{M_N + M_{N+1}} \prod_{\nu=1}^{N} \frac{M_\nu}{M_{\nu-1} + M_\nu} \frac{1}{2U_2}, \qquad (162)$$

a result evaluated at the point of entry into the last medium with the electric vector perpendicular to the plane of incidence. Computation of the complex reflectance and transmittance is relatively simple when the coefficients U_0 and U_2 of the composite quaternion U of the multilayer have been calculated.

21.5.11 Application to the monolayer. The uninitiated reader will find it a useful exercise to verify from equations (158) and (159) that for the monolayer (the case $N = 1$)

$$\begin{aligned} U_0 &= \frac{W_2}{2} e^{i\frac{\beta_1}{2}} + \frac{W_1}{2} e^{-i\frac{\beta_1}{2}}; \\ -U_2 &= \frac{1}{2} e^{-i\frac{\beta_1}{2}} + \frac{W_1 W_2}{2} e^{i\frac{\beta_1}{2}}. \end{aligned} \qquad (163)$$

It is then shown easily from equations (161), (162) and (163) that

$$\rho_0 = \frac{W_2 e^{i\beta_1} + W_1}{1 + W_1 W_2 e^{i\beta_1}}; \qquad (164)$$

$$\frac{\tau_{N+1}}{\tau_0} = \frac{\tau_2}{\tau_0} = \frac{4 M_0 M_1}{(M_0 + M_1)(M_1 + M_2)} \frac{e^{i\frac{\beta_1}{2}}}{1 + W_1 W_2 e^{i\beta_1}}. \qquad (165)$$

Equation (164) agrees, for example, with the result of equations (50) and (51) for cases $N = 1$. Likewise, equation (165) agrees with the result of equation (54). Matrix methods and quaternion methods do not possess advantages over recursion methods such as those of Sections 21.2.8 and 21.2.9 until the number N of layers in the multilayer exceeds 2. In fact, the methods of equations (164) and (165) are to be preferred as regards simplicity and convenience for the monolayer.

21.5.12 Comments. In the interesting method constructed by Gordon L. Walker, the quaternion corresponding to S_0 of equations (145) is rendered incomplete in that the coefficients of the unit vectors σ_1 and σ_2 are zero. The method presented here does not involve more quaternion multiplication than the method due to Walker and requires slightly less algebra at the last steps for computing the complex reflectance ρ_0 and the complex transmittance τ_{N+1}/τ_0 of the multilayer. Furthermore, Walker has preferred to apply the quaternion method to factored matrices $\mathcal{M}_\nu$ of the type described by equations (136). For example with respect to the matrices of equations (133) and (136), one can take

$$F = \frac{\tau_{N+1}}{2^{N+1}} \prod_{\nu=0}^{N} \frac{1}{M_\nu}; \qquad (166)$$

$$A = M_N - M_{N+1}; \quad B = M_N + M_{N+1}; \qquad (166a)$$

and compute the quaternion

$$U = \prod_{\nu=1}^{N} \tilde{\mathcal{S}}_\nu \tilde{\mathcal{J}}_\nu S_0, \qquad (166b)$$

wherein $\tilde{\mathcal{S}}_\nu$ and $\tilde{\mathcal{J}}_\nu$ are the quaternions corresponding to matrices $\mathcal{S}_\nu$ and $\mathcal{J}_\nu$, respectively, of equation (136a) and S_0 is determined from equation (145a). The complex reflectance ρ_0 can be computed from equation (161) and the complex transmittance τ_{N+1}/τ_0 can be computed from equation (160). It will be found from equations (136) and (144b) that the quaternions $\tilde{\mathcal{S}}_\nu$ and $\tilde{\mathcal{J}}_\nu$ are incomplete. The product $\tilde{\mathcal{S}}_\nu \tilde{\mathcal{J}}_\nu$ is, however, a complete quaternion.

21.6 MONOLAYER COATINGS

21.6.1 Introduction. Monolayer coatings serve many specialized purposes. One broad class of monolayers consists of dielectric substances deposited or formed upon absorbing or non-absorbing substrates. A second broad class consists of metallic or semiconducting films on absorbing or non-absorbing substrates. We shall not be concerned with that class of monolayers whose sole function is to alter the mechanical, chemical or electrical properties of a surface. Monolayers may occur naturally as, for example, in the tarnishing of silver by a layer of silver sulphide. The practical range in thickness of a layer may vary from molecular dimensions to centimeters or even meters depending mainly upon the wavelength of the radiation involved. This radiation may extend from the ultraviolet into radar. Not all monolayers should be regarded as homogeneous but a large group of monolayers can be considered homogeneous in constructing an approximate theory for interpreting their "optical" behavior. The approximations thus afforded are often in excellent agreement with experiment.

21.6.2 Methods of computation.

21.6.2.1 Wherever automatic calculators have been programmed on the basis of matrix or quaternion methods, this program can serve for computing monolayers although such programs become unduly elaborate. The following method is one of the useful and accurate methods for desk calculators. This method will be presented for the case in which the incident electric vector is perpendicular to the plane of incidence. When the electric vector vibrates in the plane of incidence, it is necessary to replace the Fresnel coefficients, W_ν, of equation (20) by the Fresnel coefficients, F_ν, of equation (21) as shown in Section 21.2.10. By setting $\nu = 1$ in equation (57) and noting that $\rho_1 = W_2$ from equation (51), one finds that the complex reflectance, ρ_o, of the monolayer is given by

$$\rho_o = \frac{W_2 \; e^{i\beta_1} + W_1}{1 + W_1 \; W_2 \; e^{i\beta_1}} , \tag{167}$$

as in equation (164). Likewise, by setting $N = 1$ in equation (54) one obtains the complex transmittance in the form

$$\frac{\tau_2}{\tau_o} = \frac{4 \; M_o \; M_1}{(M_o + M_1)(M_1 + M_2)} \; \frac{e^{i\frac{\beta_1}{2}}}{1 + W_1 \; W_2 \; e^{i\beta_1}} , \tag{168}$$

as in equation (165). Explicitly,

$$W_\nu = \frac{M_{\nu-1} - M_\nu}{M_{\nu-1} + M_\nu} \; ; \; \nu = 1, \; 2; \tag{169}$$

$$\beta_1 = \frac{4\pi}{\lambda} \; M_1 \; d_1 , \tag{169a}$$

in which M_ν is defined by equation (19). ρ_o and τ_2 are evaluated by equations (167) and (168) at the left hand side of the first interface and at the right hand side of the second interface, respectively, as indicated in Figure 21.10. Marked simplification occurs at normal incidence for which $i_o = 0$ so that $p_o = \sin i_o = 0$.

$$M_\nu = m_\nu = n_\nu \; (1 + i \, K_\nu) \; ; \; i_o = 0 . \tag{169b}$$

Furthermore, at normal incidence no distinction need be made among the directions of polarization of the electric vector when the film and substrate are isotropic. Equations (167) and (168) then serve for all states of polarization.

21.6.2.2 Phase changes that occur on reflection and transmission are given as phase retardations by arg (ρ_o) and arg (τ_2/τ_o), respectively. When these quantities are wanted, ρ_o and τ_2/τ_o must be evaluated as complex numbers by the operations indicated by equations (167) and (168).

21.6.2.3 The parameter β_1 is awkward to handle whenever m_1 is complex $(K_1 \neq 0)$. This awkwardness is increased when the angle of incidence $i_o \neq 0$. At normal incidence

$$\beta_1 = \frac{4\pi}{\lambda} \; n_1 \; d_1 \; (1 + i \, K_1) , \tag{170}$$

and

$$e^{i\beta_1} = e^{-\frac{4\pi}{\lambda} n_1 K_1 d_1} \; e^{i\frac{4\pi}{\lambda} n_1 d_1} , \tag{170a}$$

in which $4\pi \, n_1 \, d_1/\lambda$ is twice the optical path of the film and the exponent, $-4\pi \, n_1 \, K_1 \, d_1/\lambda$, is an attenuation

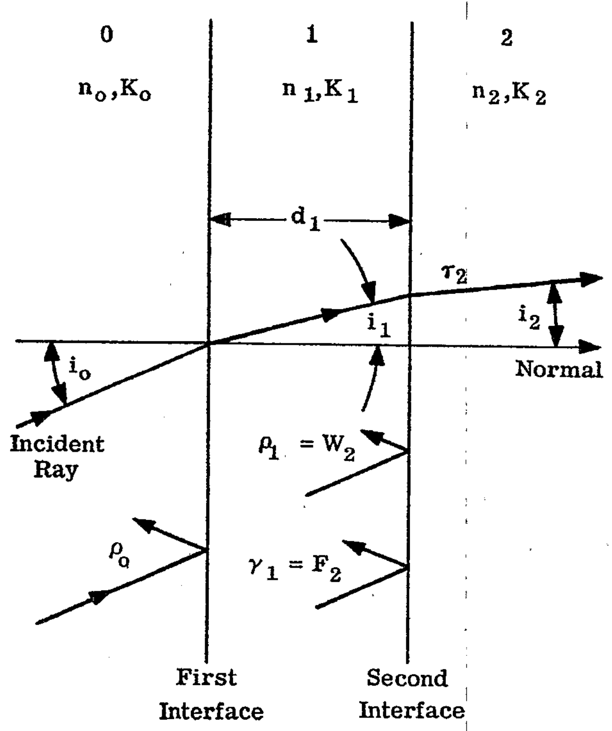


Figure 21. 10- Convention and notation with respect to the monolayer (case N = 1).

factor. When the medium of incidence and the monolayer do not absorb, and when $n_1^2 > n_o^2 p_o^2$,

$$M_1 = \left[n_1^2 - n_o^2 p_o^2\right]^{1/2} = n_1 \cos i_1 ,$$

where i_1 is the angle of refraction in the film. Correspondingly,

$$\beta_1 = \frac{4\pi}{\lambda} n_1 d_1 \cos i_1 ; \quad K_o = K_1 = 0. \tag{170b}$$

21.6.2.4 The energy reflectance R of the monolayer is given by $R = |\rho_o|^2$ and is most easily computed from equation (167) in the form

$$R = |\rho_o|^2 = \frac{|W_2 \, e^{i\beta_1} + W_1|^2}{|1 + W_1 W_2 \, e^{i\beta_1}|^2} . \tag{171}$$

Suppose that the system is non-absorbing. In such cases (a class of wide interest) the Fresnel coefficients W_1 and W_2 are real and β_1 is given by equation (170b). Correspondingly,

$$R = \frac{W_1^2 + 2 W_1 W_2 \cos \beta_1 + W_2^2}{1 + 2 W_1 W_2 \cos \beta_1 + W_1^2 W_2^2} . \tag{172}$$

Differentiation of R with respect to β_1 in equation (172) shows that the extreme values of $R(\beta_1)$ occur when

$$\beta_1 = \nu \pi ; \quad \nu \text{ an integer.} \tag{173}$$

Hence when the electric vector is perpendicular to the plane of incidence, when $n_1 > n_o p_o$ and when $K_o = K_1 = K_2 = 0$, the maxima and minima, R_m, are given by

$$R_m = \frac{(W_1 \pm W_2)^2}{(1 \pm W_1 W_2)^2} . \tag{174}$$

The values of W_1 and W_2 (and hence R_m) depend upon the angle of incidence. But at fixed angles of incidence, $R_{max.}$ and $R_{min.}$ remain fixed and R is a periodic function of the thickness d_1. Departures from periodicity can occur when β_1 is altered by changing wavelength because the refractive indices (and hence W_1 and W_2) are dispersive with λ. It is not difficult to see that $R(\beta_1)$ cannot be periodic when the film absorbs. Equation (170a) shows that the $\exp(i\beta_1)$ approaches zero as d_1 approaches infinity. Consequently, from equation (171) R oscillates with increasing β_1 such that

$$R \rightarrow |W_1|^2 , \tag{175}$$

the Fresnel coefficient of reflectance of the first interface, Figure 21.10.

21.6.2.5 When the system contains no absorbing media, the energy transmittance, T_e, of the monolayer can be found in terms of the energy reflectance, R, from the law of conservation of energy, namely,

$$T_e + R = 1 , \tag{176}$$

irrespective of the refractive indices of the first and last media. When absorption occurs, one can compute the complex transmittance τ_2 / τ_o from equation (168) and the energy transmittance from equation (58). If only the film is absorbing, $(n_a)_2 = n_2$ so that equation (58) yields the result,

$$T_e = \frac{n_2 \cos i_2}{n_o \cos i_o} \left| \frac{\tau_2}{\tau_o} \right|^2 , \tag{177}$$

for cases in which the electric vector is perpendicular* to the plane of incidence.

21.6.3 Non-absorbing systems; normal incidence.

21.6.3.1 The behavior of the energy reflectances, $|\rho_o|^2$, for non-absorbing monolayers on non-absorbing substrates is illustrated in Figure 21.11 for cases in which the refractive index $n_o = 1$. When no absorption occurs, reversal of the direction of incidence leaves $|\rho_o|^2$ unchanged. For reasons explained in Section 21.2.14, the monolayer is an absentee layer at points $\beta_1 = \nu 2\pi$ where $\nu = 0$ or an integer. At these points the energy reflectance is that of the uncoated glass.

21.6.3.2 With respect to cases $n_o < n_1 < n_2$, it follows from equation (169) that $W_1 < 0$ and $W_2 < 0$. Correspondingly from equation (174),

$$R_m = R_{min.} ,$$

when

$$\beta_1 = \mu \pi ; \quad \mu \text{ an odd integer.} \tag{178}$$

By introducing

$$W_1 = \frac{n_o - n_1}{n_o + n_1} ; \quad W_2 = \frac{n_1 - n_2}{n_1 + n_2}$$

into equation (174), one finds that

$$R_{min.} = |\rho_o|^2_{min.} = \left(\frac{n_o n_2 - n_1^2}{n_o n_2 + n_1^2} \right)^2 . \tag{179}$$

In particular, $R_{min.} = 0$ when β_1 obeys equation (178) and n_1 is chosen so that

$$n_1 = \sqrt{n_o n_2} . \tag{180}$$

Equations (178) and (180) agree with the more general equation (115) and (116) at normal incidence. Equation (179) is important for two reasons. First, it enables one to estimate the minimum value of the energy reflectance. Secondly, it enables one to compute the refractive index n_1 of the film from the measured** value of $R_{min.}$ and the known values of n_o and n_2. The refractive indices of evaporated monolayers are usually less than those of the bulk materials and vary with the conditions of evaporation.

---

* See equations (70) and (71) for cases in which the electric vector vibrates in the plane of incidence.

**The spectral energy reflectance, R_p, of a coated plate is ordinarily obtained with a spectrophotometer. Suppose, for example, that the absorption of the plate is negligible and that the energy reflectance of the back surface is B. On account

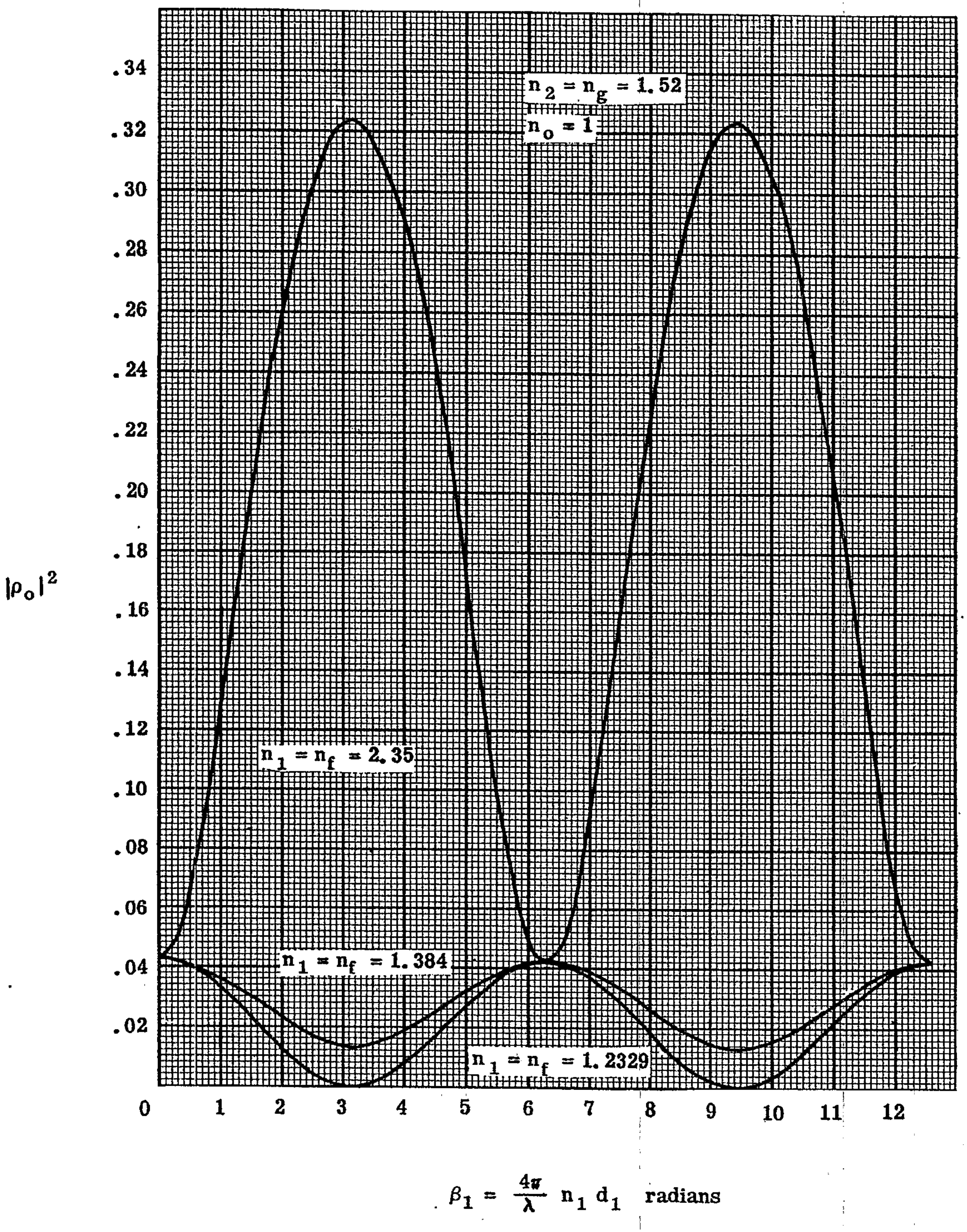


Figure 21. 11- Energy reflectances $|\rho_o|^2$ vs β_1 in radians. These curves illustrate the periodic behavior of nonabsorbing monolayers on nonabsorbing substrates. β_1 is now twice the optical path of the monolayer. The curves drawn for $n_f = 2.35$ and $n_f = 1.384$ with $n_g = 1.52$ correspond to zinc sulphide and magnesium fluoride, respectively, on spectacle crown glass. The curves illustrate the important characteristic that $|\rho_o|^2 \gtrless$ reflectance of the uncoated substrate according as $n_f \gtrless n_g$ at points $\beta_1 \neq \nu\pi$ where $\nu = 0$ or an integer. The curve for $n_f = 1.2329 = \sqrt{n_g} = \sqrt{1.52}$ illustrates how zero reflectance is achieved at points $\beta_1 = \mu\pi$ where μ is an odd integer.

of the interreflections that occur within the plate, it follows that

$$R_p = \frac{R + B - 2RB}{1 - RB} , \tag{1}$$

where R is the surface reflectance $R = \left| \rho_o \right|^2$. Hence

$$R = \left| \rho_o \right|^2 = \frac{R_p - B}{1 + R_p B - 2B} . \tag{2}$$

If the plate is a glass plate in air with its back surface uncoated, $B = \left[(n_g - 1)/(n_g + 1) \right]^2$ where n_g is the refractive index of the glass plate.

21.6.3.3 When $n_o < n_1 > n_2$, equation (169) shows that Fresnel's coefficients W_1 and W_2 have opposite sign. At points $\beta_1 = \mu \pi$ (μ odd), the extreme reflectances R_m of equation (174) are now maxima with the choice of the negative sign. Also $|\rho_o|^2_{max.}$ is therefore given by the right hand member of equation (179). $|\rho_o|^2_{max.}$ approaches unity with increasing n_1 .

21.6.3.4 In Figure 21.12, $|\rho_o|^2$ is plotted against λ in the visible region for the indicated values of $n_1 = n_f$ and $n_2 = n_g$ with $n_o = 1$. The family of curves illustrates the effectiveness of various monolayers in reducing surface reflectances of spectacle crown glass. The thicknesses of the monolayers are chosen so that the optical path of each monolayer is one-fourth wavelength at 0.550 microns.

21.6.3.5 Energy reflectances $|\rho_o|^2$ are plotted against wavelength in the infrared region in Figure 21.13 for $n_o = 1$, $n_2 = n_s = 3.450$ and for the indicated values of $n_1 = n_f$. This family of curves illustrates the effectiveness of various monolayers in reducing the reflectance of a surface of silicon. As applied to silicon, the effects of absorption by the substrate are not included in Figure 21.13. Absorption of silicon is low in the infrared region. The curve for $n_f = 1.52$ has been added to illustrate what occurs when n_f is smaller than the value required for producing zero reflectance, i.e. when $n_f < \sqrt{n_o n_2}$. Let n_f and n_f' denote the refractive indices of two different non-absorbing monolayers such that $n_f > \sqrt{n_o n_2}$ and $n_f' < \sqrt{n_o n_2}$. It is not difficult to show that when the refractive indices are independent of wavelength, the two monolayers produce the same spectral reflectance curves provided that n_f and n_f' are chosen in accordance with the relation

$$n_f \; n_f' = n_1 \; n_1' = n_o \; n_2 . \tag{181}$$

For example, a monolayer having the refractive index $n_f' = 1.57$ is equivalent to the monolayer having the refractive index $n_f = 2.200$ when n_o and n_2 have been chosen as in Figure 21.13.

21.6.4 Dielectric monolayers on opaque, metallic substrates; normal incidence. Dielectric layers of hard materials such as magnesium fluoride and silicon monoxide are often deposited upon surfaces of aluminum, silver and other metals for the purpose of protecting these surfaces from abrasion. The effects of monolayers of $M_g F_2$ and SiO upon the surface reflectance and phase change on reflection are illustrated in Figure 21.14 for opaque substrates of aluminum. Appreciable losses in reflectance can occur when the monolayers are so thin that their optical paths $n_1 d_1 < \lambda/4$. For maximum reflectance, the optical path of the monolayer should be slightly less than $\lambda/2$. This maximum reflectance can exceed that of the uncoated surface provided that the refractive index n_1 of the monolayer is high enough. With $M_g F_2$(7) and SiO,(8) one obtains a maximum reflectance that departs imperceptibly from that of the uncoated substrate. The phase change on reflection varies markedly with the optical path of the monolayer. This property is of importance to interferometry.

21.6.5 Absorbing monolayers on opaque, metallic substrates; normal incidence. The effects of absorbing monolayers upon the surface reflectance of opaque, metallic substrates are illustrated in Figure 21.15 by a series of monolayers that have a fixed and relatively high refractive index $n_1 = 4.0$. The curve for the non-absorbing film $n_1 K_1 = 0$ shows that the maximum reflectance can exceed * that of the uncoated surface appreciably when n_1 becomes high. Increasing the absorption of the monolayer within the range of $n_1 K_1$ of Figure 21.15 reduces both the maximum and the minimum reflectances. The minimum reflectances occur for optical paths p_1 near 45° or $\lambda/8$. The family of curves suggest that tarnishing of silver is due to the formation of a monolayer.

---

(7) Monolayers of $M_g F_2$ have become valuable for increasing the reflectance of aluminum in the ultraviolet. For a discussion of the region around 1216A, see P.H. Berning, G. Hass and R. P. Madden; Paper T51 given during 44th Annual Meeting of OSA, Ottawa, October 1959.

(8) The refractive indices and absorption of so called SiO films depend markedly upon conditions of evaporation. Slow deposition in the presence of air or oxygen produces "SiO" films that can have refractive indices near 1.7 with low absorption. See G. Hass, Vacuum, 2, p 338 (1952).

* Compare with Figure 21.14

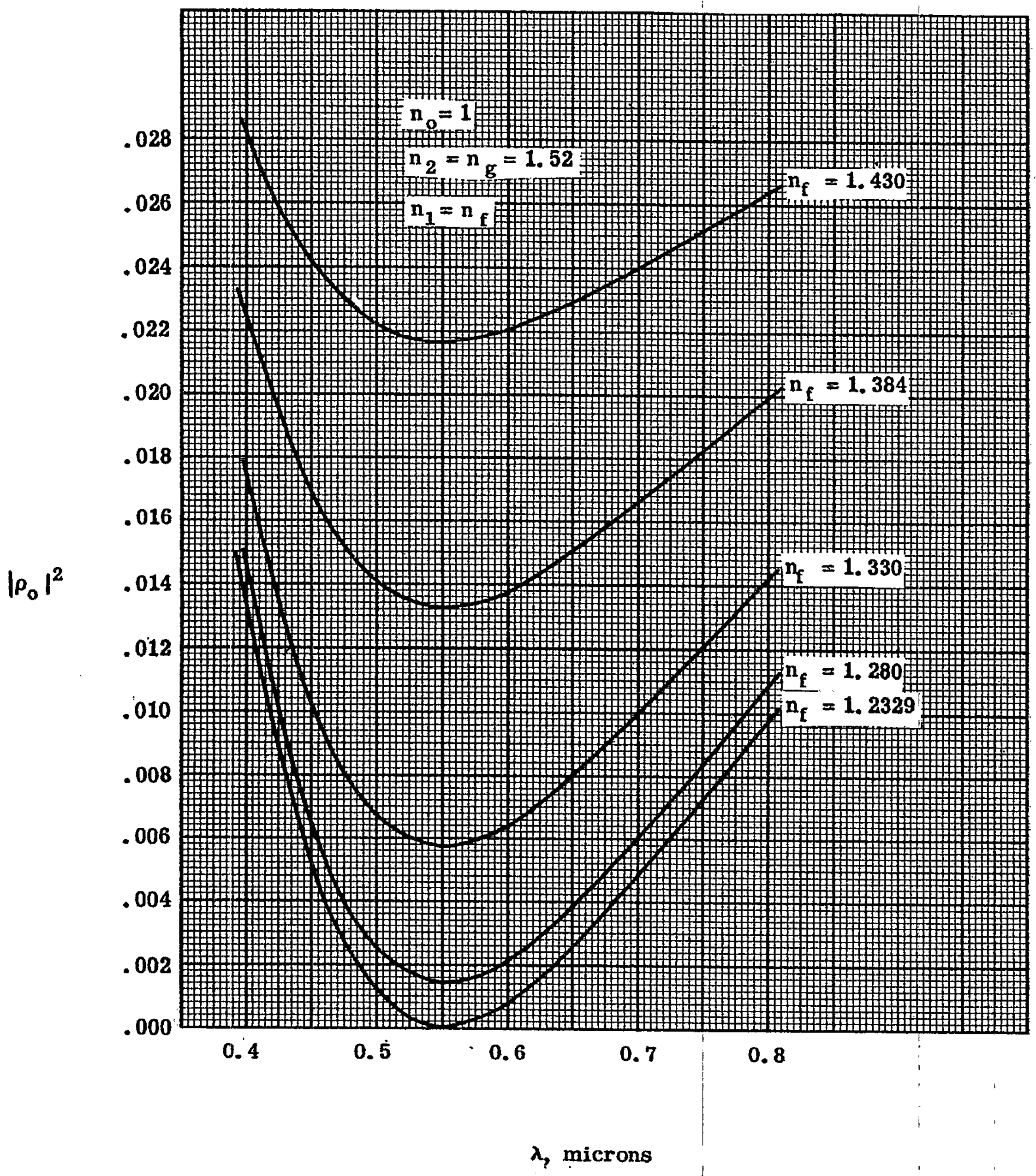


Figure 21. 12- Plot of energy reflectances vs wavelength in microns for various low reflecting monolayers on spectacle crown glass. Each monolayer has the optical path $\lambda/4$ at $\lambda = 0.550$ microns.

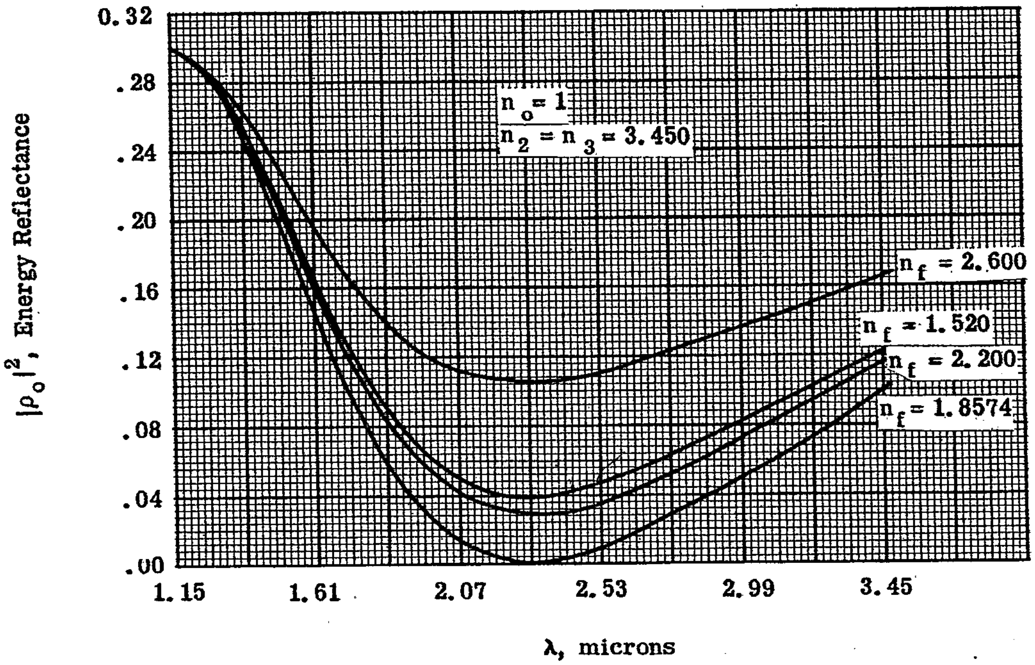


Figure 21. 13- Low reflecting monolayers on substrates having high refractive index.

21.6.6 Metallic monolayers on glass; normal incidence. The curves of Figure 21.16 a and b illustrate the optical behavior of monolayers of silver on non-absorbing substrates such as glass. Layers of silver differ from layers of other metals mainly in that layers of silver have remarkably low * absorption although the nK -value is moderately high. Silver is sensibly opaque of thicknesses near 0.15λ. It should be observed that the reflectance at the glass to film interface is less than that at the air to film interface -- a characteristic of most metals. As the thickness of the silver layer is increased, the phase changes on reflection pass into the third quadrant for reasons discussed following equation (16) in Section 21. 2. 3. The reflectance curve for reflection from glass to silver exhibits theoretically a minimum at a thickness between zero and 0.01λ.

21.6.7 Non-absorbing systems; oblique incidence. The manner in which low reflecting, non-absorbing monolayers modify the reflectances $|\rho_o|^2$ and $|\gamma_o|^2$ of surfaces of non-absorbing substrates is illustrated** by Figures 21.17, 21.18, and 21.19 for the 45° angle of incidence. The spectral reflectance curves of Figure 21.17 for $|\rho_o|^2$ and $|\gamma_o|^2$ correspond to electric vectors that vibrate respectively perpendicular and parallel to the plane of incidence for a monolayer of $M_g F_2$ on a surface of spectacle crown glass. Whereas a monolayer of $M_g F_2$ is effective in reducing both $|\rho_o|^2$ and $|\gamma_o|^2$ for spectacle crown glass, the reflectance $|\rho_o|^2$ is still quite high. Figure 21.18 shows that the refractive index of the monolayer must approach the value $n_1 = 1.2$ in order to reduce $|\rho_o|^2$ below 1% for spectacle crown glass. Rugged films having refractive indices below 1.30 are not available. On the other hand, Figure 21.19 shows that one can choose a relatively high and available value of n_1 for reducing both $|\rho_o|^2$ and $|\gamma_o|^2$ to values below 1.2% when the refractive index n_2 of the substrate is markedly higher than that of spectacle crown glass. Figure 21.19 illustrates also the fact that $|\rho_o|^2_{min.} = |\gamma_o|^2_{min.}$, when $n_1 = \sqrt{n_o n_2}$. The effect of increasing or decreasing the angle of incidence from 45° is to raise or to lower, respectively, the reflectance at the crossing point C of Figure 21.19. It is to be expected that the reduction of $|\rho_o|^2$ presents the more formidable problem; for $|\gamma_o|^2$ is automatically zero at Brewster's angle of incidence.

---

*The amount of absorption is altered considerably by the conditions of evaporation. L. G. Schultz gives the values $n = 0.055$ and $nK = 3.32$ for silver at $\mu = 0.55$.

**As with many other subjects of this text, the possible number of instructive illustrations is restricted in the interests of conserving space.

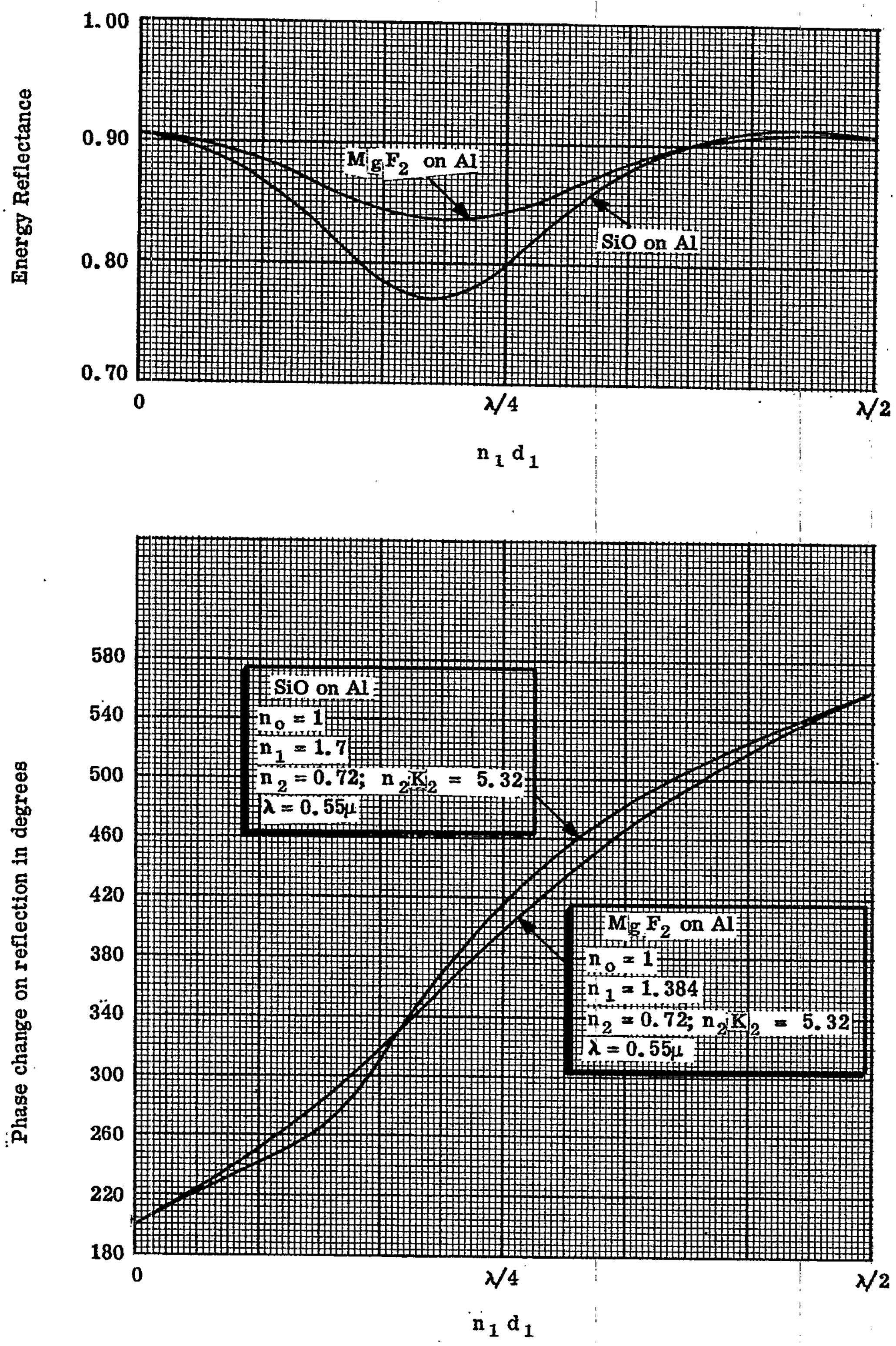


Figure 21. 14- Energy reflectances and phase changes on reflection vs optical path of the film in wavelengths for M_gF_2 and SiO on opaque substrates of aluminum. Phase changes on reflection appear as phase retardations. Absorption by SiO monolayers has been neglected. The optical constants of aluminum are those given by L. G. Schultz, J. Opt. Soc. Amer., 44, 357-368 (1954).

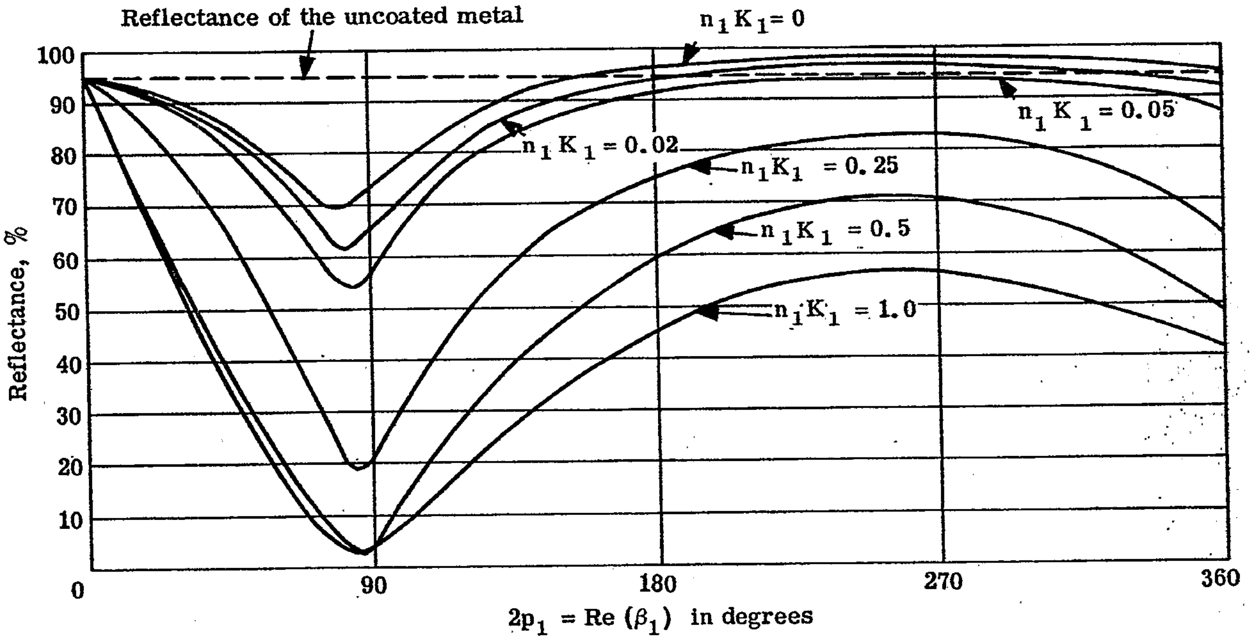


Figure 21. 15- Reflectances from various absorbing monolayers on a metallic substrate as functions of $R_e(\beta_1) = 2p_1$ where p_1 is the optical path through the monolayer. The metallic substrate corresponds to silver with the optical constants $n_2 = 0.15$ and $n_2K_2 = 3.28$. The monolayers have the high refractive index $n_1 = 4.0$ and the indicated values of n_1K_1. The case $n_1K_1 = 0$ is the nonabsorbing monolayer.

21. 7 BILAYER COATINGS

21. 7. 1 Introduction. Bilayers possess advantages over monolayers for decreasing or increasing the reflectance of surfaces of dielectrics or metals. As examples, zero reflectance can be obtained at an assigned wavelength by suitable choice of the thickness ratio of the two layers and a marked degree of control over the distribution of spectral reflectance or transmittance becomes possible. Bilayers are superior to monolayers for beam splitters and for control of phase changes that occur on reflection or transmission. Whereas absorbing bilayers are used for such purposes as controlling the transmittance of sunglasses, the most important bilayers are predominantly dielectric. Low reflecting bilayers fall into well defined groups whose characteristics will be discussed.

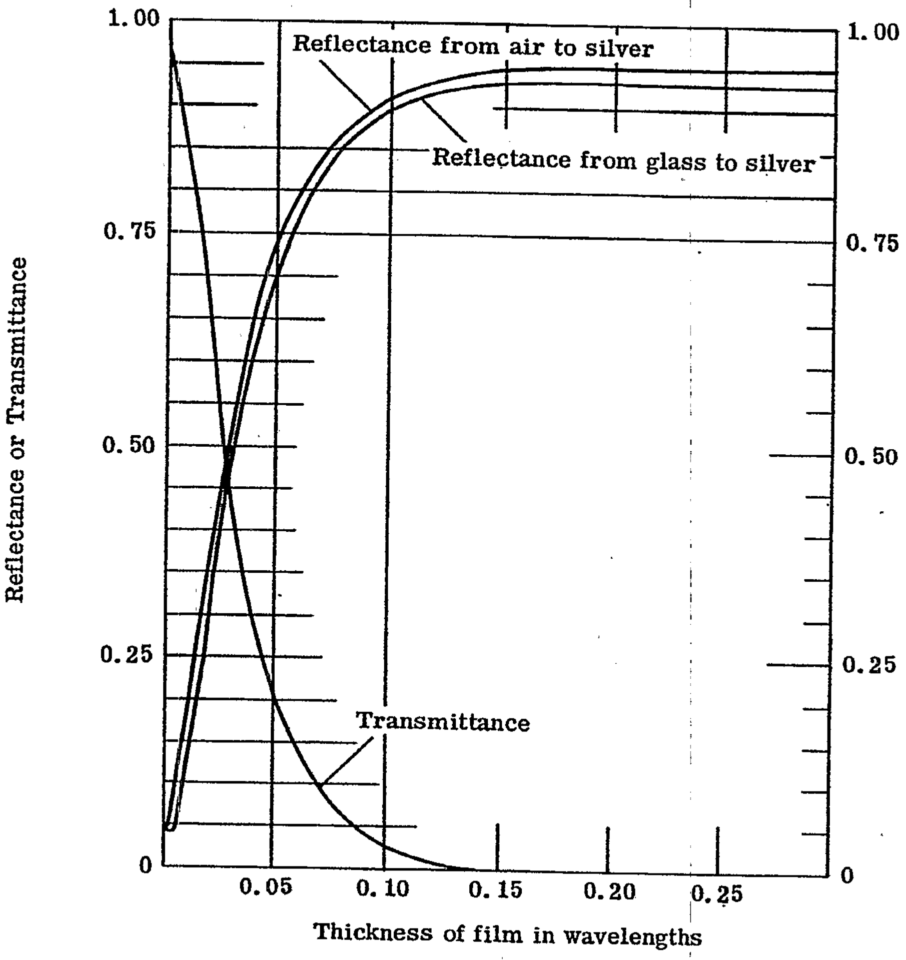
21. 7. 2 Methods of computation. Matrix and quaternion methods have been treated in Sections 21. 4 and 21. 5. The reader who is inclined toward watching the interfacial reflectances and transmittances and toward using recursion formulae may choose the method of Sections 21. 2. 8 to 21. 2. 11. The convention and notation with respect to bilayers is shown in Figure 21. 20. Suppose that the electric vector is perpendicular to the plane of incidence. We obtain the complex reflectance, ρ_o, in the form

$$\rho_o = \frac{\rho_1 e^{i\beta_1} + W_1}{1 + W_1 W_2 e^{i\beta_1}}, \tag{182}$$

wherein

$$\rho_1 = \frac{W_3 e^{i\beta_2} + W_2}{1 + W_2 W_3 e^{i\beta_2}}. \tag{182a}$$

The Fresnel coefficients of reflection W_ν are defined, as usual, by equations (19) and (20). β_ν $(\nu = 1, 2)$ is defined by equation (56). The complex transmittance, τ_3, corresponding with ρ_o is obtained by setting $N = 2$, $\tau_o = 1$ and $\rho_2 = W_3$ in equation (54). The complex reflectance, γ_o, for cases in which the electric vector vibrates in the plane of incidence can be computed from equation (182), after replacing ρ_ν by γ_ν and W_ν by F_ν as defined by equation (21). The corresponding complex transmittance is denoted by T_3 and is obtained by setting $N = 2$, $T_o = 1$ and $\gamma_2 = F_3$ in equation (70). At normal incidence, $\gamma_o = \rho_o$ and $T_3 = \tau_3$. Closed formulae are available for the energy reflectances. These formulae simplify markedly



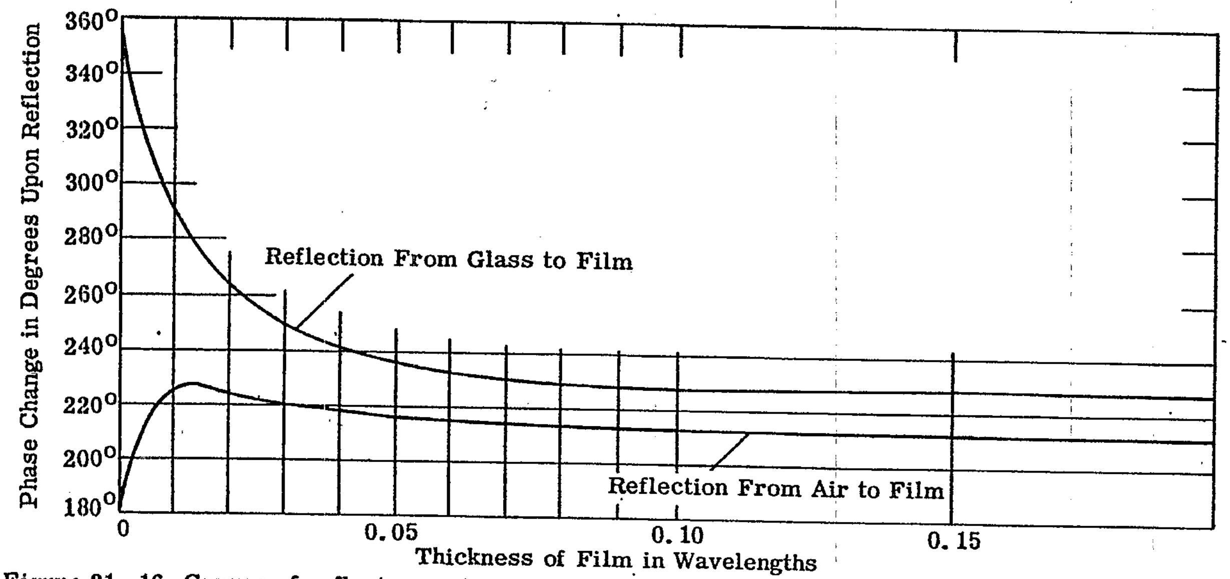


Figure 21. 16- Curves of reflectance, transmittance and phase changes on reflection vs thickness of the monolayer in wavelengths for silver films on spectacle crown glass. These curves have been computed for glass having the refractive index 1.52 and for an absorbing film having the optical constants $n_f = 0.15$ and $n_f K_f = 3.28$. These optical constants predate those measured for silver by L. G. Schultz and belong to a wavelength near 0.55 microns. The phase changes of Figure 21.16b are retardations. Phase retardations of 230° can be regarded, if desired, as phase advances of 130°.

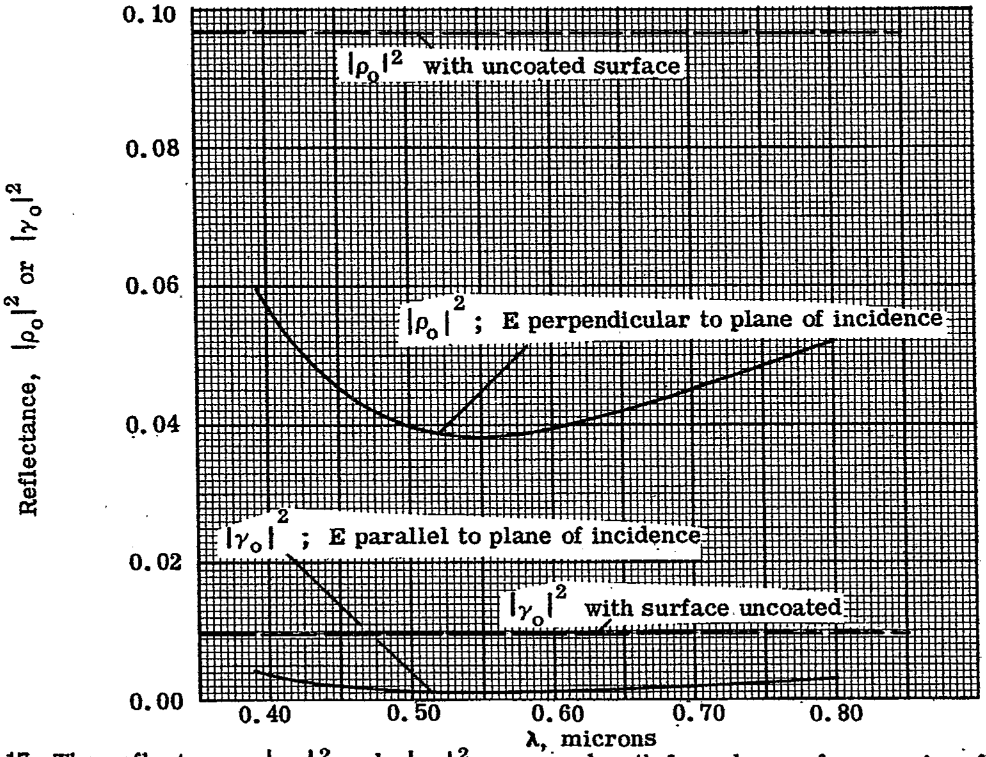


Figure 21. 17- The reflectances $|\rho_o|^2$ and $|\gamma_o|^2$ vs wavelength for a layer of magnesium fluoride on spectacle crown glass. The refractive indices are $n_o = 1$; $n_1 = 1.384$ and $n_2 = 1.52$ and the angle of incidence is 45°. The thickness of the film has been chosen so that $\beta_1 = 4\pi n_1 d_1 \cos i_1/\lambda = \pi$ at $\lambda = 0.55\mu$.

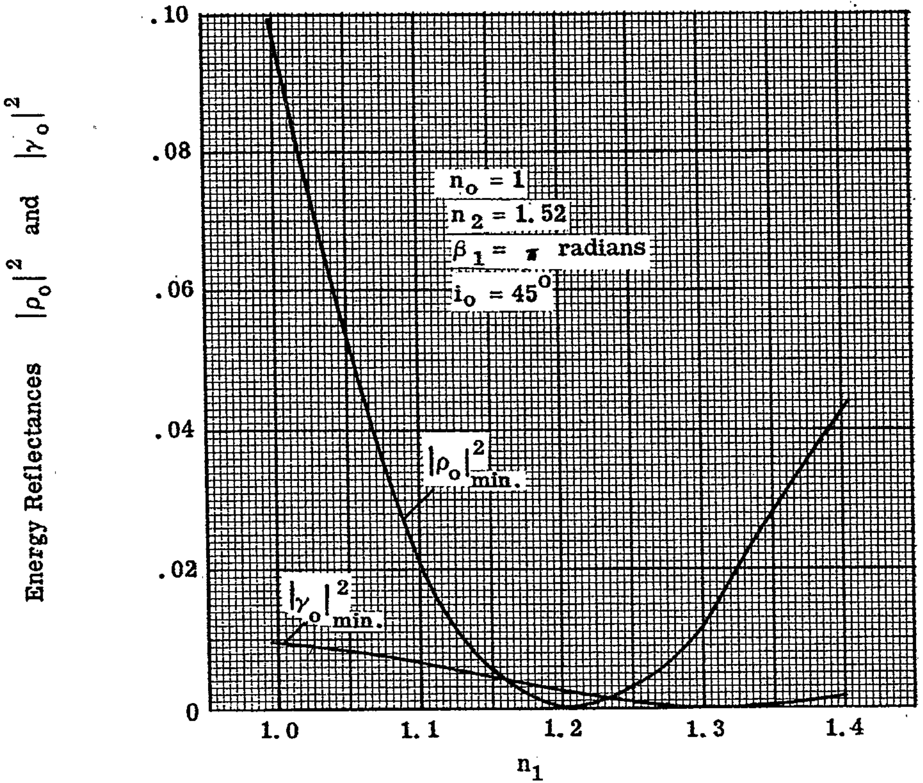


Figure 21. 18- Plot of the minimum reflectances $|\rho_o|^2_{min.}$ and $|\gamma_o|^2_{min.}$ against the refractive indices n_1 of the monolayers on spectacle crown glass at 45° angle of incidence.

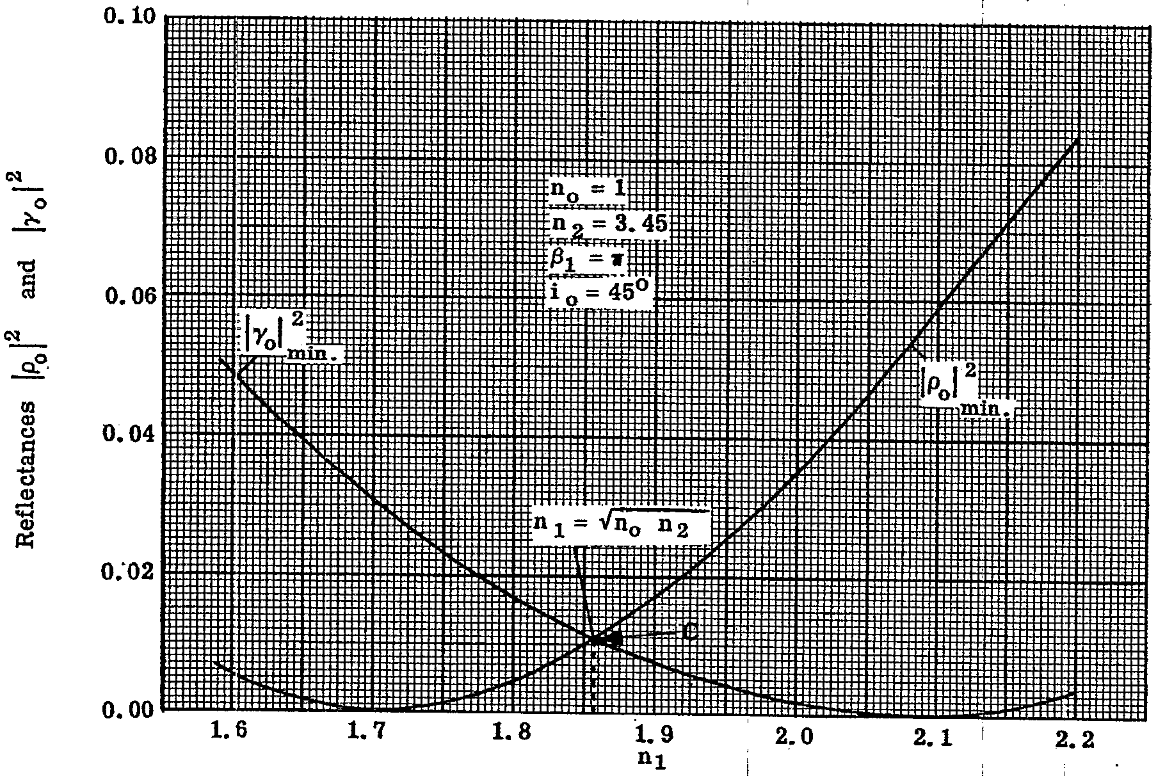


Figure 21. 19- Plot of the minimum reflectances $|\rho_o|^2_{min.}$ and $|\gamma_o|^2_{min.}$ against n_1 for monolayers on a nonabsorbing substrate of high refractive index $n = 3.45$ at the fixed angle 45^o of incidence.

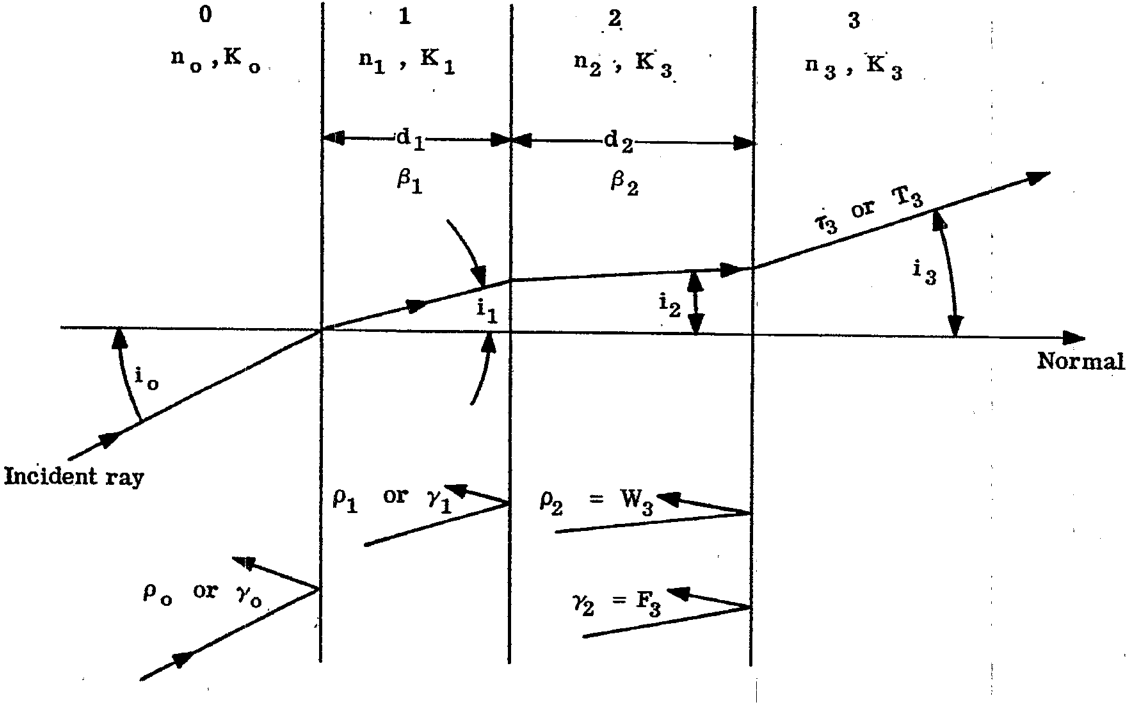


Figure 21. 20- Notation and convention with respect to bilayers (case $N = 2$).

only when all of the media and layers are non-absorbing and when total internal reflectance need not be considered. Thus,

$$|\rho_o|^2 = 1 - \frac{(1-W_1^2)(1-W_2^2)(1-W_3^2)}{D}, \tag{183}$$

where

$$D = 1 + W_1^2 W_2^2 + W_1^2 W_3^2 + W_2^2 W_3^2 + 2 W_1 W_2 (1 + W_3^2) \cos \beta_1$$
$$+ 2 W_2 W_3 (1 + W_1^2) \cos \beta_2 + 2 W_1 W_3 \cos(\beta_1 + \beta_2)$$
$$+ 2 W_1 W_3 W_2^2 \cos(\beta_1 - \beta_2) . \tag{183a}$$

W_ν and β_ν are now real with

$$\beta_\nu = \frac{4\pi}{\lambda} n_\nu d_\nu \cos i_\nu ; \quad \nu = 1, 2. \tag{183b}$$

Equation (183) can be used to compute $|\gamma_o|^2$ by replacing ρ_o by γ_o and W_ν by F_ν. Because absorption has been excluded, the sum of the energy reflectance and the energy transmittance must be unity when

$$|\tau_o| = |T_o| = 1 .$$

21.7.3 The simplest low-reflecting bilayer. Monolayers such as $M_g F_2$ on glass fail to produce zero reflectance because $n_1 > \sqrt{n_o n_2}$. This class of monolayers is characterized by the fact that $n_o < n_1 < n_2$. Consequently, $W_1 < 0$ and $W_2 < 0$ at normal incidence. The following conclusions are not difficult to ascertain from equation (169) for normal incidence.

$$|W_2| < |W_1| \quad \text{when} \quad n_1 > \sqrt{n_o n_2} ; \tag{184}$$

$$|W_2| > |W_1| \quad \text{when} \quad n_1 < \sqrt{n_o n_2} . \tag{184a}$$

Hence the absolute value of the Fresnel coefficient W_2 at the film-to-glass interface is too small to satisfy the condition for zero reflectance when $n_1 > \sqrt{n_o n_2}$; for $W_1 = W_2$ at the point of zero reflectance. These considerations suggest that it should be possible to decrease the energy reflectance from monolayers for which $n_1 > \sqrt{n_o n_2}$ by depositing either a thin dielectric film that has high refractive index or a thin film of highly reflecting metal upon the interface between media no. 1 and 2 as illustrated in Figure 21.21. The author(9) has shown that metals can be used to augment the interfacial reflectance for obtaining zero reflectance. An advantage of this method is that a wide range of dielectric materials can be used as the monolayer by making suitable choices of the thickness of the thin metallic film. A disadvantage of employing metallic or absorbing films for augmenting the interfacial reflectance is that the energy reflectance for incidence from glass to film can be quite high even when the energy reflectance in the opposite direction from air to film has been made zero. This disadvantage as regards the irreversibility of the reflectance can be avoided by using dielectric materials of high refractive index as the reflectance augmenting layer in the manner described by Osterberg, Kashdan and Pride.(10) The use of dielectrics having high refractive index instead of metal yielded some unexpected advantages that will now be discussed.

21.7.3.1 Summary of the theory. In summarizing the results of an unpublished theory of the author, let us utilize the convention of Figure 21.22 in which layer no. 1 is to be the thin dielectric layer of high refractive index. All of the media are non-absorbing and the incidence is to be normal. One can show from the zero condition ($\rho_o = 0$) of equation (98) that the value of β_1 required for zero reflectance is given by

$$\cos \beta_1 = \frac{W_1^2 + W_2^2 - W_3^2 (1 + W_1^2 W_2^2)}{2 W_1 W_2 (W_3^2 - 1)}, \tag{185}$$

in which

$$\beta_1 = \frac{4\pi}{\lambda} n_1 d_1 \text{ radians.} \tag{185a}$$

---

(9) U. S. Patent 2,366,687 (Jan. 2, 1945).

(10) See abstracts of papers published in J. Opt. Soc. Amer., 42, p. 291 (1952).

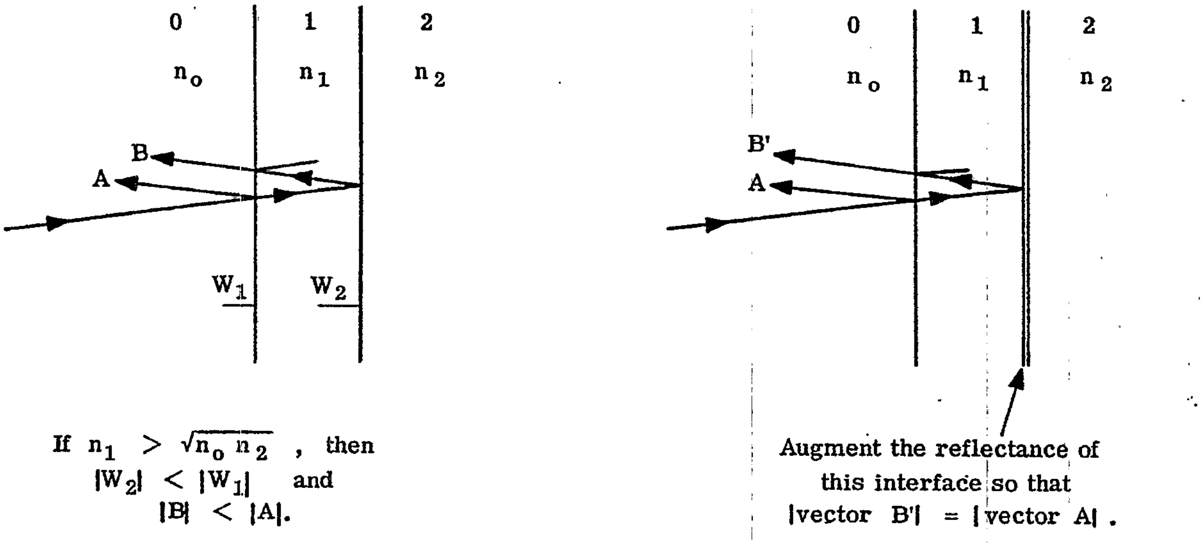


Figure 21.21- Illustration of the principle of augmenting one of the interfacial reflectances of a monolayer so as to obtain zero or decreased reflectance by depositing a thin, highly reflecting layer upon the interface.

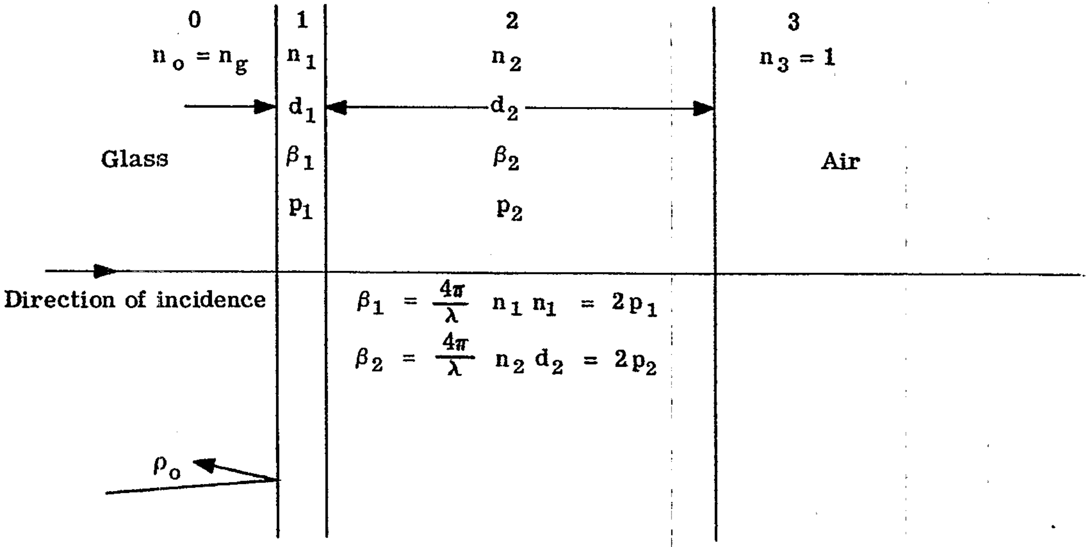


Figure 21.22- Notation with respect to the simplest low reflecting bilayer. Layer no. 1 is a thin film of high refractive index n_1 . p_1 and p_2 are the optical paths of the two layers.

Correspondingly, in terms of the refractive indices

$$\cos \beta_1 = 1 + \frac{2n_1^2 (n_g - 1)\ (n_g - n_2^2)}{(n_1^2 - n_g^2)\ (n_1^2 - n_2^2)}, \qquad (185b)$$

for cases in which $n_3 = n_{air} = 1$. For solutions to exist, it is necessary that

$$\cos \beta_1 \leqq 1, \qquad (185c)$$

a condition that places restrictions upon the refractive indices of equation (185b). Of these restrictions, the one of greatest interest is

$$n_1^2 > n_2^2 \ n_g > \left\{ \begin{matrix} n_2^2 \\ n_g^2 \end{matrix} \right. > n_g . \qquad (185d)$$

If equation (185b) has a solution $\beta_1 = \beta_{10}$, it has other solutions that differ from β_{10} by integral multiples of 2π. We are to choose the smallest possible solution for which

$$0 < \beta_1 < \pi . \qquad (185e)$$

Correspondingly, the optical path p_1 of the thin layer will be less than $\pi/2$. It can be much less than $\pi/2$, as we shall see. Having computed β_1, one can compute the associated value of β_2 from β_1 in the following manner. Let

$$\theta_1 = \tan^{-1} \frac{-W_1 \ \sin \beta_1}{-(W_1 \ \cos \beta_1 + W_2)} ; \qquad (186)$$

$$\theta_2 = \tan^{-1} \frac{W_1 \ W_2 \ \sin \beta_1}{1 + W_1 \ W_2 \ \cos \beta_1} . \qquad (186a)$$

Then

$$- \beta_2 = \theta_1 - \theta_2 ; \quad (\theta_2 > \theta_1). \qquad (186b)$$

There exists no ambiguity as to the quadrant in which θ_1 and θ_2 fall because

$$\begin{aligned} \sin \theta_1 &: - W_1 \ \sin \beta_1 ; & \sin \theta_2 &: W_1 \ W_2 \ \sin \beta_1 ; \\ \cos \theta_1 &: - (W_1 \ \cos \beta_1 + W_2); & \cos \theta_2 &: 1 + W_1 \ W_2 \ \cos \beta_1 ; \end{aligned} \qquad (186c)$$

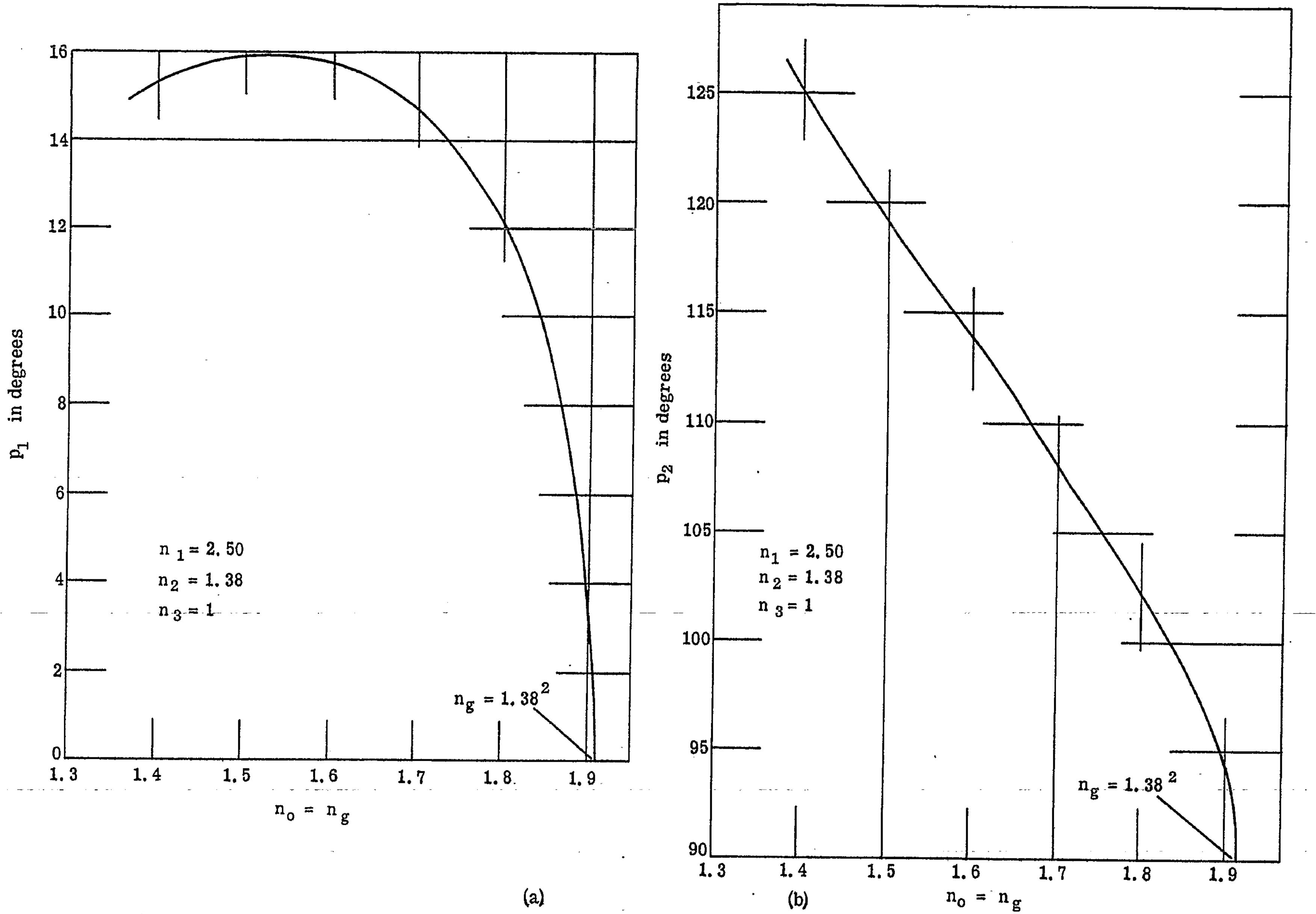
in which the factor of proportionality is greater than zero. Equations (185) and (186) thus enable one to compute * the doubled optical paths β_1 and β_2, Figure 21.12, of the two layers when n_1 has been suitably chosen with respect to n_g and n_2.

21.7.3.2 Let us now consider the case: $n_1 = 2.50$; $n_2 = 1.38$; $n_3 = 1$. This case applies to bilayers comprised of a thin inner layer of TiO_2 and an outer layer of $M_g F_2$ under conditions at which the refractive indices of TiO_2 and $M_g F_2$ are 2.50 and 1.38, respectively. The optical paths p_1 and p_2 required for vanishing reflectance have been computed from equations (185) and (186) and plotted as functions of n_g in Figure 21.23 (a) and (b). Curiously, the required optical path p_1 of the inner layer with refractive index $n_1 = 2.50$ is substantially 15.5° or 0.043λ for the range n_g of ordinary glasses. This fact has been confirmed experimentally and means that a TiO_2 layer of fixed thickness will serve for practically all glasses. This thickness is only $0.043/2.5 = 0.0172\lambda$. The required thickness approaches zero as n_g approaches $n_2^2 = 1.38^2$. The corresponding optical path p_2 of the outer layer must be chosen with some care since the slope of the curve of Figure 21.23(b) is quite marked. The optical path, p_2, exceeds the quarter wave condition considerably in the range, n_g, for ordinary glasses. However, the main effect of altering the optical path of the outer layer is to alter the wavelength at which minimum reflectance occurs. This class of bilayer is therefore relatively easy to produce. It is extremely durable. Cerium oxide is to be preferred to titanium dioxide as the inner layer because cerium oxide can be evaporated as a dielectric material without need for subsequent heating and oxidation. Equations (185) and (186) should be applied to redetermine p_1 and p_2 when cerium oxide is used. Comparisons of the bilayer (theoretical and experimental) with a monolayer of

---

*Frequently, one can be quite certain about the quadrant in which β_2 must fall. Under such circumstances it is simpler to compute β_2 from the formula

$$\cos \beta_2 = \frac{W_2^2 + W_3^2 - W_1^2 \ (1 + W_2^2 \ W_3^2)}{2 W_2 \ W_3 \ (W_1^2 - 1)} .$$



(a) (b)

Figure 21.23- Plot of the optical paths p_1 and p_2 against the refractive index n_g of the substrate. p_1 and p_2 are the optical paths required to obtain zero reflectance for the layer of high refractive index $n_1 = 2.50$ and the layer of low refractive index n_2, respectively.

$M_g F_2$ on spectacle crown glass of refractive index $n_g = 1.52$ are made in Figure 21.24. The theoretical and experimental curves for the bilayer are in good agreement and exhibit much lower reflectances than monolayers of $M_g F_2$ near the wavelength for minimum reflectance. The dispersion of the reflectance is, however, markedly greater for the bilayer.

21.7.3.3 With reference to white light reflectance of the bilayer, T. Sawaki and H. Kubota investigated bilayers having low white light reflectance and found that those bilayers which simulate monolayers having augmented interfacial reflectance rank among the bilayers exhibiting the lowest white light reflectances.

21.7.4 Achromatized bilayers.

21.7.4.1 Curves of the computed * spectral reflectances at normal incidence for various achromatized bilayers on spectacle crown glass are illustrated in Figures 21.25 and 21.26. The outer layer is exemplified by $M_g F_2$. It has a refractive index n_1 which fails to meet the zero condition, $n_1 = \sqrt{n_o n_2}$, for monolayers. The outer and inner layers have the optical path $\lambda/4$ and $\lambda/2$, respectively, at a chosen ** wavelength, λ_m. The inner layer of refractive index n_2 is therefore an absentee layer at $\lambda = \lambda_m$. Consequently, the reflectance R_m at $\lambda = \lambda_m$ is due to the outer layer and the substrate of refractive index $n_3 = n_g$. When $n_2 > n_1$, the reflectance R_m is a maximum for wavelengths near λ_m. Achromatization occurs because the energy reflectance must drop before it can rise as λ is varied on either side of λ_m. With respect to the curves of Figure 21.25, the minimum reflectances on each side of point R_m are equal. Bilayers exhibiting this property will be classified as isoachromatic. The two minima of the spectral reflectance curve for the case $n_1 = 1.38$, $n_2 = 1.86$ and $n_2 = 1.52$ are zeros. Isoachromatic bilayers displaying zero minima will be called null-isoachromatic bilayers. The publications by A. F. Turner (11) and A. Vasicek (12) deal with null-isoachromatic bilayers.

21.7.4.2 The spectral reflectance curve for $n_2 = 1.52$, i.e. for the comparison monolayer of $M_g F_2$, of Figures 21.25 and 21.27 lies above the other spectral reflectance curves except at extreme values of λ and β_1. To substitute any one of the isoachromatic bilayers of Figures 21.25 and 21.27 in place of the monolayer of refractive index n_1 will therefore reduce the white light reflectance. The class of bilayer discussed in Section 10.7.3 and its subsections is, however, much superior for reducing white light reflectance. Except for specialized applications, achromatic bilayers are to be preferred when increased neutrality of spectral reflectance becomes important. For this purpose, the null-isoachromatic bilayers are not as suitable as the isoachromatic bilayers exemplified by the curve for $n_2 = 1.55$ of Figure 21.27. The flatness of this curve is quite remarkable. Comparison of Figures 21.25 and 21.26 shows that a wide range of distributions of spectral reflectances can be attained by means of achromatized bilayers. Figure 21.28 has been included to show how the separation of the minima at points A and B can be reduced by choosing n_1 nearer to the value $n_1 = \sqrt{n_o n_g}$.

21.7.4.3 The theoretical design of null-achromatic bilayers will now be developed algebraically in a manner that illustrates one use of the method of admittances. From equations (81a) and (81b), one obtains for normal incidence the results,

$$Y_3 = -n_3; \tag{187}$$

$$Y_1 = n_1 \frac{Y_2 - in_1 \tan p_1}{n_1 - in_2 \tan p_1}; \quad p_1 = \frac{\beta_1}{2}; \tag{187a}$$

$$Y_2 = n_2 \frac{Y_3 - in_2 \tan p_2}{n_2 - iY_3 \tan p_2}; \quad p_2 = \frac{\beta_2}{2}. \tag{187b}$$

The condition for zero reflectance is $\rho_o = 0$ or, from equation (79a),

$$Y_1 = -M_o = -n_o = -1 \tag{188}$$

since we shall suppose that the medium of incidence is vacuum. By eliminating the admittances Y_1, Y_2 and Y_3, one obtains quite directly the condition for zero reflectance in its general form for non-absorbing bilayers, namely,

$$n_1 \frac{1 - in_1 \tan p_1}{n_1 - i \tan p_1} = n_2 \frac{n_3 + in_2 \tan p_2}{n_2 + in_3 \tan p_2}; \tag{188a}$$

---

* The materials of this discussion have been taken from unpublished research notes of the author.

** λ_m is often chosen as 0.55μ when the bilayer is intended for the visible region.

(11) A. F. Turner, J. de Phys., 11, 444 (1950).

(12) A. Vasicek, Optica Acta, May 1951, special issue, pp. 20-25.

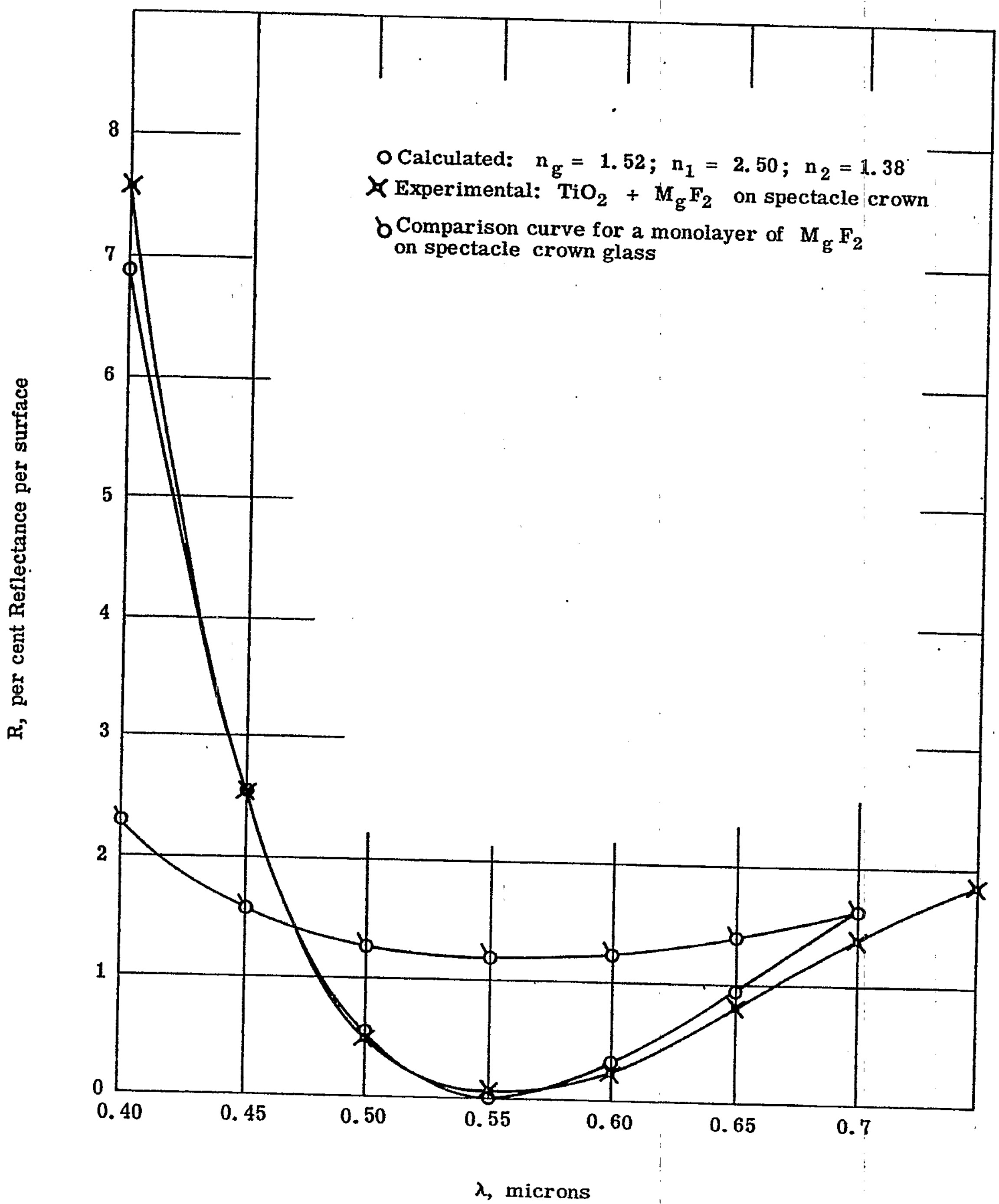


Figure 21.24- Comparison of the computed and experimental reflectances from bilayers of $TiO_2 + MgF_2$ on spectacle crown glass of refractive index 1.52. In making the computations it has been assumed that the refractive indices remain fixed at the values $n_g = 1.52$; $n_1 = 2.50$ and $n_2 = 1.38$. A spectral reflectance curve for monolayers of MgF_2 on spectacle crown glass is included for further comparison.

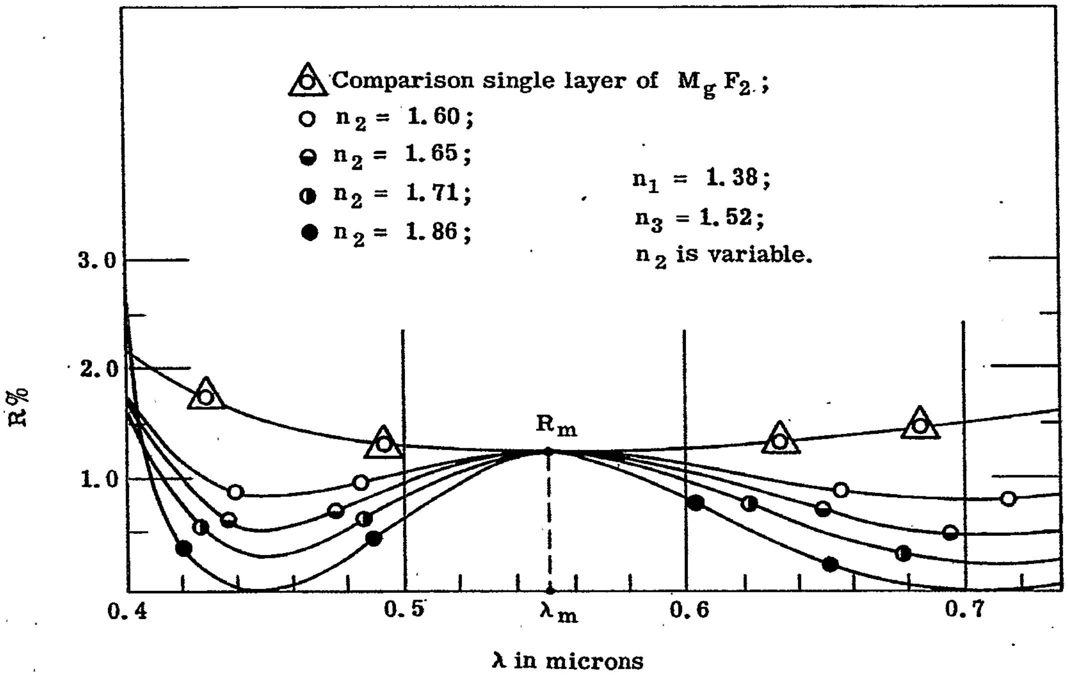


Figure 21.25 - Spectral reflectance curves for a family of isoachromatic bilayers on spectacle crown glass of refractive index 1.52.

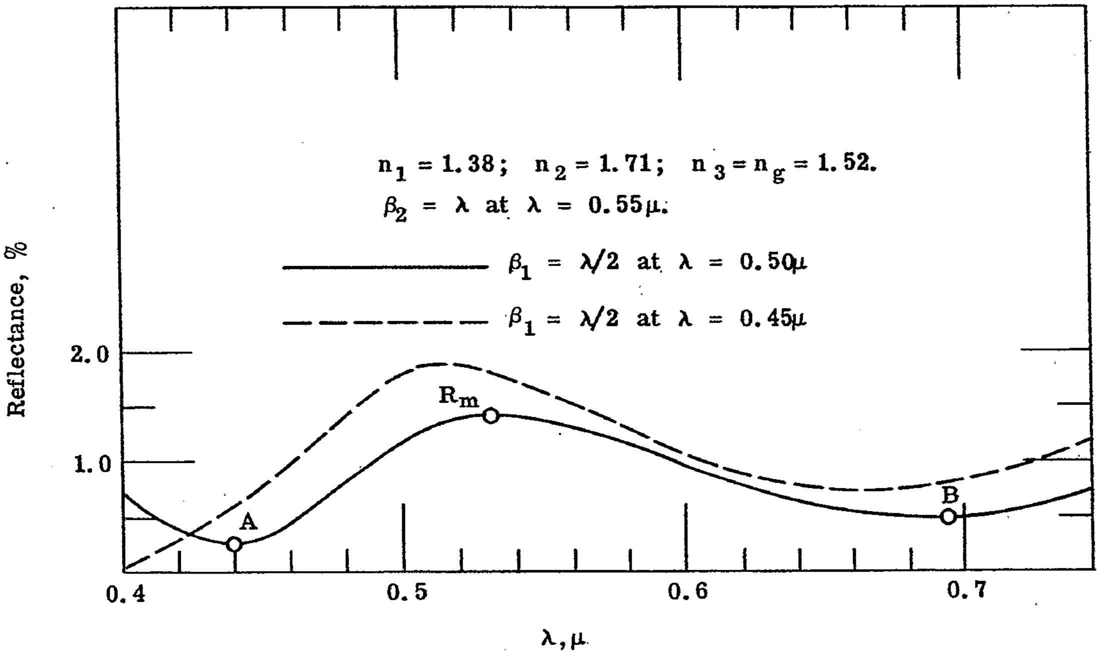


Figure 21.26 - Spectral reflectance curves for two achromatic but not isoachromatic bilayers on spectacle crown glass. These curves illustrate the trend wherein the minimum reflectances at A and B become unlike when the outer layer of low refractive index and the inner layer of higher refractive index are not quarter wave and half wave, respectively, in optical path at the same wavelength.

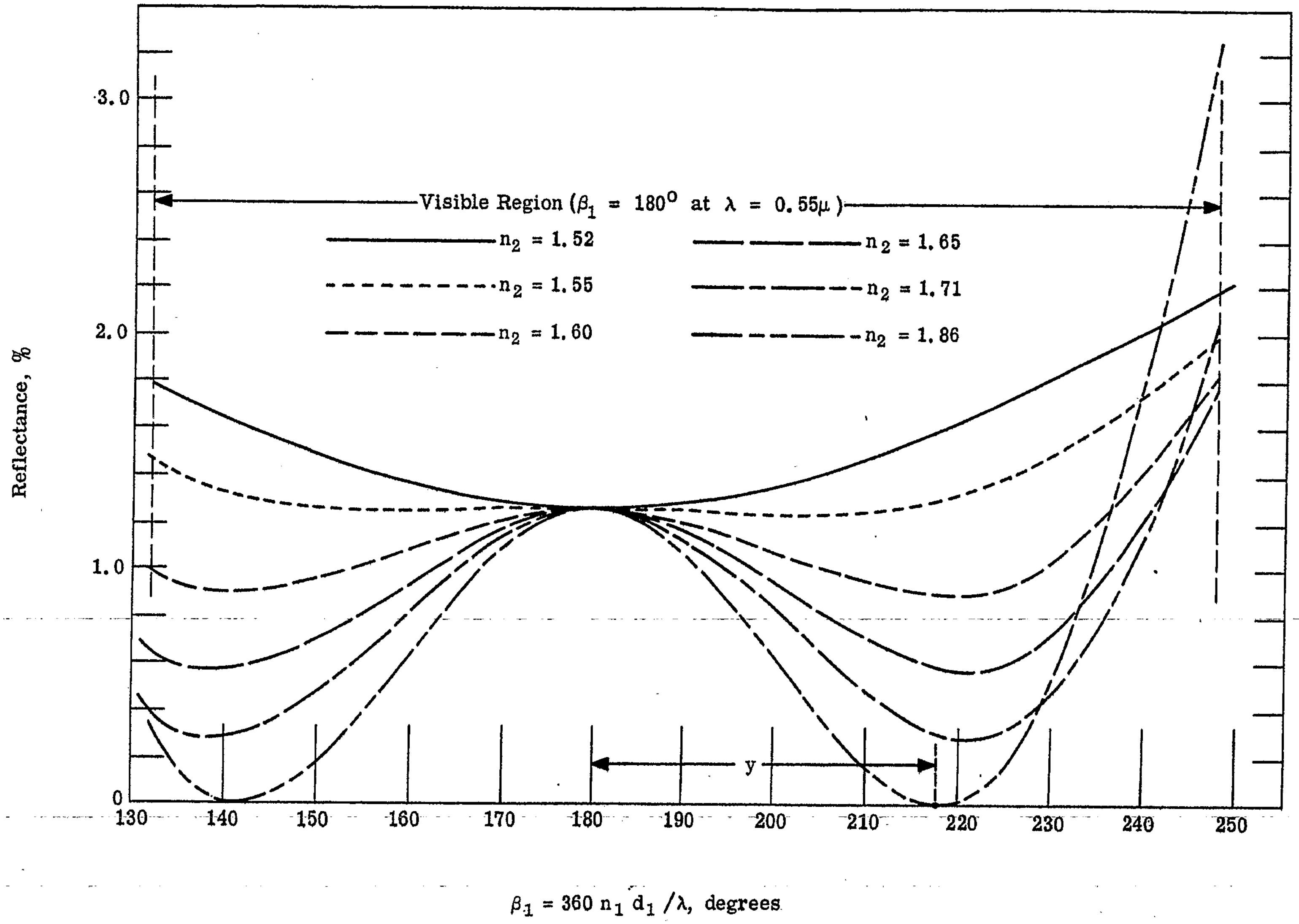


Figure 21.27 - Curves illustrating the symmetry of the energy reflectances of isoachromatic bilayers about the point $\beta_1 = 180^\circ$ where β_1 is twice the optical path of the outer layer of low refractive index $n_1 = 1.38$ (corresponding to MgF_2). The refractive index of the substrate is 1.52 as in Table 21.14.

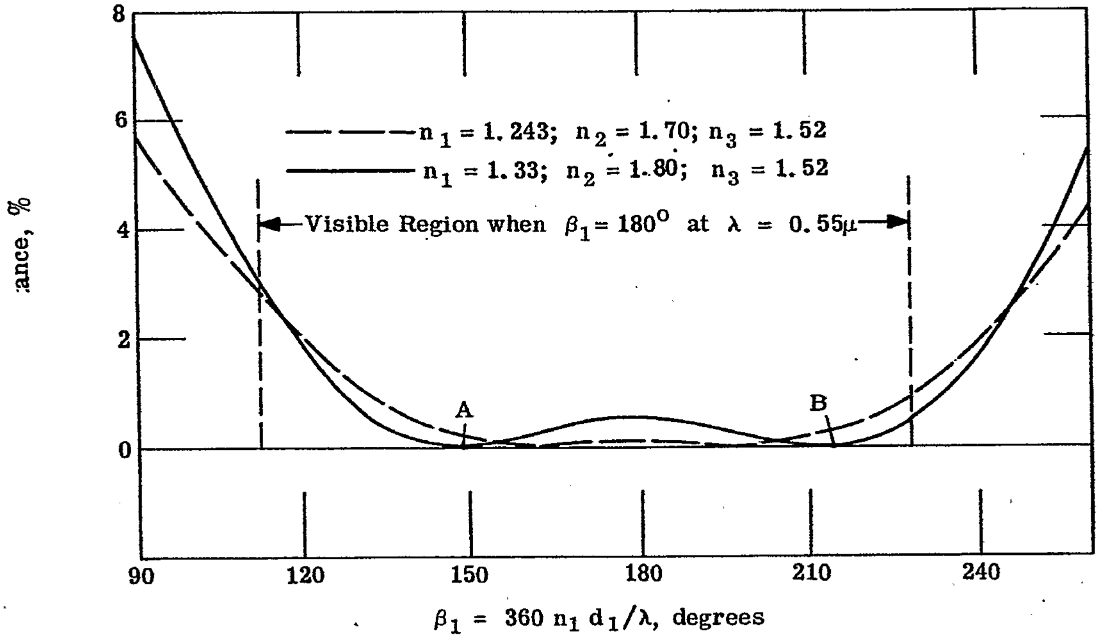


Figure 21.28- Curves showing the effects of reducing the refractive index n_1 of the outer layer at fixed values of n_3.

in which

$$p_1 = 2\pi n_1 d_1/\lambda ; \quad p_2 = 2\pi n_2 d_2/\lambda . \tag{188b}$$

We suppose at this point that the ratio n_2/n_1 does not depend upon wavelength and introduce

$$p_2 = f \, p_1 , \tag{188c}$$

in which f can be assigned any desired value. Investigation shows that null-isoachromatic bilayers are possible mathematically for the choice $f = 1$, but require values of n_1 that are usually too small to obtain physically. The choice $f = 2$ leads to the null-isoachromatic bilayers that are of interest to this section. Equating real and imaginary parts on each side of equation (188a), after clearing fractions and introducing

$$p_2 = 2 p_1 = 2 p , \tag{189}$$

one obtains the zero condition in the form

$$n_1 (1 - n_3) + \left[\frac{2n_1^2 n}{n_2} - 2n_2 - n_1(1 - n_3)\right] \tan^2 p = 0; \tag{189a}$$

$$\frac{2n_1 \; n_3}{n_2} - 2n_1 n_2 + n_3 - n_1^2 - (n_3^2 - n_1^2) \tan^2 p = 0; \tag{189b}$$

By eliminating $\tan^2 p$, one obtains the result

$$n_2 n_3 (n_1^2 + 1) (n_1 + n_2) - 2n_1 (n_2^3 + n_1 n_3^2) = 0, \tag{190}$$

an equation for determining n_2 from n_1 and n_3 or n_2 from n_1 and n_3. One obtains also

$$\cos 2p = \cos \beta_1 = n_1 \frac{n_2^2 - n_3}{n_1 (n_3 - n_2^2) + n_2 (n_3 - n_1^2)} . \tag{190a}$$

Introduce

$$2p = \beta_1 = 180^{\circ} \pm y \,, \tag{190b}$$

where y is indicated in Figure 21.27. Then the solution for y becomes

$$\cos y = n_1 \frac{n_2^2 - n_3}{n_1(n_2^2 - n_3) + n_2(n_1^2 - n_3)} \,. \tag{190c}$$

The required refractive indices and the optical paths, p_1 and p_2, of the two members of the bilayer thus become known algebraically.

21.7.5 Quarter wave bilayers.

21.7.5.1 Non-absorbing, quarter wave bilayers on non-absorbing substrates comprise a third, important class of bilayers. In the interests of brevity, the following discussion will be restricted to normal incidence.

21.7.5.2 Let us suppose that layer number two, Figure 21.20, is a quarter wave layer. Then $\beta_2 = \pi$ and ρ_1, equation (182a), will be real for non-absorbing systems. Consequently from equation (182),

$$|\rho_o|^2 = \frac{\rho_1^2 + W_1^2 + 2\rho_1 W_1 \cos\beta_1}{1 + \rho_1^2 W_1^2 + 2\rho_1 W_1^2 \cos\beta_1} \,. \tag{191}$$

By differentiating $|\rho_o|^2$ with respect to β_1, one finds that the condition for maxima and minima is $\sin\beta_1 = 0$ or $\beta_1 = \nu\pi$ where ν is an integer. The choice $\nu = 1$ makes layer number one a quarter wave layer. If, therefore, both members of the bilayer are quarter wave layers at a given wavelength, λ_o, then the energy reflectance $|\rho_o|^2$ is either a minimum or a maximum for wavelengths in the immediate neighborhood of λ_o. Maxima and minima can occur at wavelengths removed from λ_o but at these wavelengths the bilayer will not be a quarter wave bilayer.

21.7.5.3 Consider now the recursion formula (81a) at wavelength λ_o for which $\beta_1 = \beta_2 = \pi$. One obtains

$$Y_{\nu-1} = \frac{M_{\nu-1}^2}{Y_\nu} = \frac{n_{\nu-1}^2}{Y_\nu} \,. \tag{192}$$

Whence

$$Y_1 = \frac{n_1^2}{Y_2} \;;\quad Y_2 = \frac{n_2^2}{Y_3} \,. \tag{192a}$$

From equation (81b) we note that $Y_3 = -n_3$. Therefore the admittance Y_1 of the quarter wave bilayer is given by

$$Y_1 = -\frac{n_1^2}{n_2^2}\, n_3 \tag{192b}$$

at $\lambda = \lambda_o$.

21.7.5.4 Equations (79a) and (192b) give the complex reflectance ρ_o of the non-absorbing bilayer on a non-absorbing substrate at $\lambda = \lambda_o$ and at normal incidence in the form

$$\rho_o = \frac{n_o - \left(\frac{n_1}{n_2}\right)^2 n_3}{n_o + \left(\frac{n_1}{n_2}\right)^2 n_3} \,, \tag{193}$$

from which the energy reflectance $|\rho_o|^2$ is either a maximum or a minimum. The condition that must exist among the refractive indices to obtain zero reflectance is

$$n_o n_2^2 = n_3 n_1^2 \,. \tag{194}$$

Available materials(13) do not ordinarily meet the zero condition of equation (194).

21.7.5.5 Let us now consider the case $n_o = 1$. In most applications, the medium of incidence is air for which n_o may be set at the approximate value unity. Then

$$R = |\rho_o|^2 = \left(\frac{1 - \left(\frac{n_1}{n_2}\right)^2 n_3}{1 + \left(\frac{n_1}{n_2}\right)^2 n_3} \right) ; \quad \lambda = \lambda_o ; \qquad (195)$$

where R denotes energy reflectance. The manner in which R depends on the choice of n_1 and n_2 is illustrated in Figure 21.29 for the case $n_3 = n_g = 1.52$. Whereas only restricted generalizations can be made about combinations n_1 and n_2 that produce reflectances less than the reflectance of the uncoated substrate, we may conclude that the reflectance of the bilayer exceeds that of the uncoated substrate whenever $n_1 > n_2$ for the case $n_o = 1$. To obtain a high reflecting bilayer, one should deposit first the layer having the lower refractive index n_2.

21.7.6 Non-quarter wave bilayers.

21.7.6.1 Quarter wave bilayers will rarely satisfy the zero condition (194). However, when equation (194) is not satisfied by the refractive indices n_1 and n_2, it may be possible to choose β_1 and β_2 different from 180° in such a manner as to meet the more general zero condition. The corresponding bilayers belong to a fourth important class. The bilayers of Section 21.7.3 modified by adding $\lambda/2$ to the optical path of the inner layer of high refractive index are examples of this fourth class of bilayers. Of greatest interest are those bilayers for which β_1 and β_2 depart only slightly from 180°. The exact method of equations (185) to (185b) and equations (186) can be applied to find members of this fourth class.

---

(13) For a discussion of methods of chemical deposition that utilize a mixture of materials having high and low refractive indices in order to obtain films having refractive indices in the approximate range 1.44 to 2.1, see U. S. Pat. 2466119, April 15, 1949 by H. R. Moulton and E. D. Tillyer.

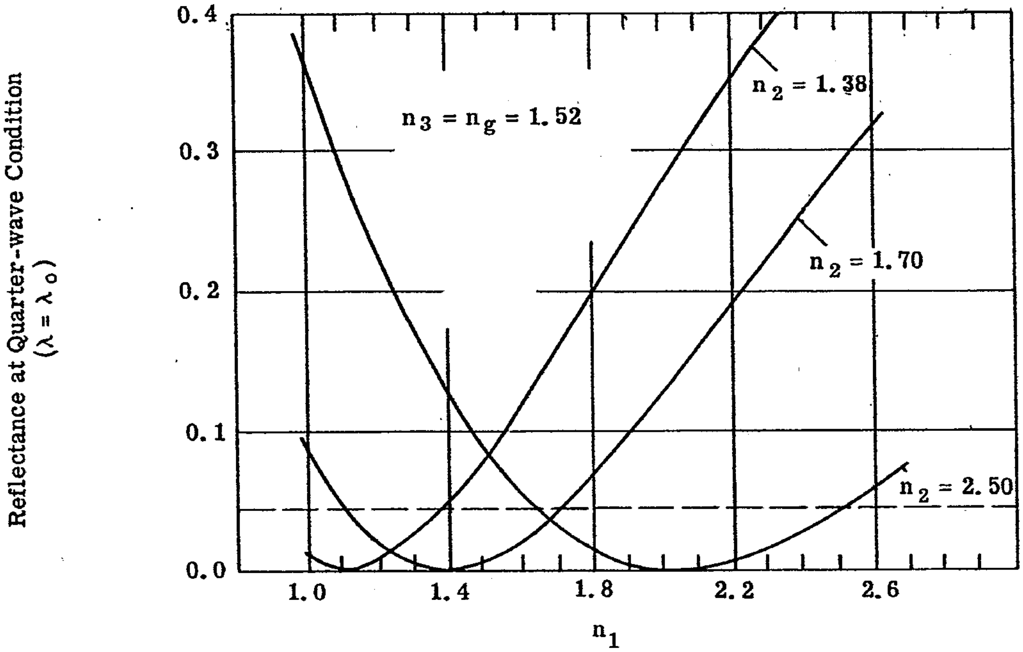


Figure 21.29- Plots of energy reflectances at $\lambda = \lambda_o$ vs the refractive index n_1 of the outer layer for the indicated values of the refractive indices n_2 of the inner layer with $n_3 = n_g = 1.52$. The broken line indicates the reflectance of the uncoated substrate, i.e. reflectance from air to the uncoated glass of refractive index $n_g = 1.52$.

21.7.6.2 The early investigators of thin films utilized a graphical method, involving closed polygons, for finding multilayers having zero reflectance. The approximate method of Section 21.2.11 is the algebraic equivalent of the method of closed polygons when applied to the zero condition of equation (98). By inserting ρ_1 from equation (75) into equation (98), one finds that

$$W_3 \; e^{i(\beta_1 + \beta_2)} + W_2 \; e^{i\beta_1} + W_1 = 0. \qquad (196)$$

By equating the real and imaginary parts of the left hand member of equation (196) to zero, one obtains the zero condition in the form

$$\sin(\beta_1 + \beta_2) = -\frac{W_2}{W_3} \sin\beta_1 \qquad (197)$$

$$W_3 \cos(\beta_1 + \beta_2) + W_2 \cos\beta_1 + W_1 = 0. \qquad (197a)$$

Elimination of $(\beta_1 + \beta_2)$ from equation (197a) with the aid of equation (197) yields straightforwardly the result

$$\cos\beta_1 = \frac{W_3^2 - W_1^2 - W_2^2}{2 \, W_1 \, W_2} \, . \qquad (197b)$$

Equations (197) and (197b) determine in a simple manner first β_1 and then β_2. As the first example, consider the case in which $n_o = 1$, $n_1 = 1.38$, $n_2 = 1.72$ and $n_3 = 1.52$. With reference to the curve for $n_2 = 1.70$, Figure 21.29, n_2 is then a little too high to obtain zero reflectance when $\beta_1 = \beta_2 = 180^{\circ}$. Equations (197b) and (197) yield four pairs of solutions for β_1 and β_2. These are

$$\beta_1 = 164^{\circ} \; 16' \; ; \; \beta_2 = \begin{cases} 224^{\circ} \; 36' \\ 346^{\circ} \; 52' \end{cases} ; \qquad (198)$$

and

$$\beta_1 = 199^{\circ} \; 44' \; ; \; \beta_2 = \begin{cases} 135^{\circ} \; 24' \\ 13^{\circ} \; 08' \end{cases} . \qquad (198a)$$

The case $\beta_1 = 199^{\circ} \; 44'$ and $\beta_2 = 13^{\circ} \; 08'$ belongs to the classification of Section 21.7.3. In case $\beta_1 = 164^{\circ} \; 16'$ and $\beta_2 = 224^{\circ} \; 36'$, the optical paths $\beta_1/2$ and $\beta_2/2$ are most nearly equal to 90°. On the other hand, when one examines the examples $n_o = 1$, $n_2 = 1.68$, $n_3 = 1.52$ with $n_1 = 1.38$ and 1.384, he finds that $|\cos\beta_1| > 1$. Although the physical parameters have been changed only slightly, the method of closed polygons does not admit solution.

21.7.7 High reflecting bilayers on metals. Bilayers of non-absorbing films can be used for gaining significant increases in reflectance from metals. For example, a very durable bilayer of silicon monoxide and titanium oxide for increasing the reflectance of an evaporated aluminum mirror has been described by G. Hass.(14)

21.8 TRILAYERS

21.8.1 Introduction. The main interest in trilayers has centered on the possibilities which they provide for obtaining lower and flatter curves of spectral reflectances than is feasible with bilayers. Trilayers can exhibit three minima in spectral reflectance over the visible region. In extreme cases all three of these minima can be zero minima. The so called quarter-half-three quarter wave trilayer (15) is advantageous for achromatization. *

21.8.2 Low reflectance trilayers. The behavior of the quarter-half-quarter wave type of low reflecting trilayer is indicated in Figure 21.30. It should be noted that the central layer is a half-wave (and hence absentee) layer at a wavelength λ_o at which the inner and outer layers are quarter wave layers. Consequently, the trilayer behaves in effect as a quarter wave bilayer at $\lambda = \lambda_o$. The condition for zero reflectance can be found by considering equation (194) for quarter wave bilayers. Thus, $n_o^2 \, n_3^2 = n_4 \, n_1^2$ when the refractive indices

---

(14) Georg Hass, Vacuum, 2, p 339 (1952).

(15) For an example and discussion of this class of trilayer see O. S. Heavens, Optical Properties of Thin Solid Films, Butterworths Scientific Publications, London (1955), pp 213-215.

* The term apochromatization would be more appropriate when the film is designed to have three minima, i.e. is "corrected" at three wavelengths.

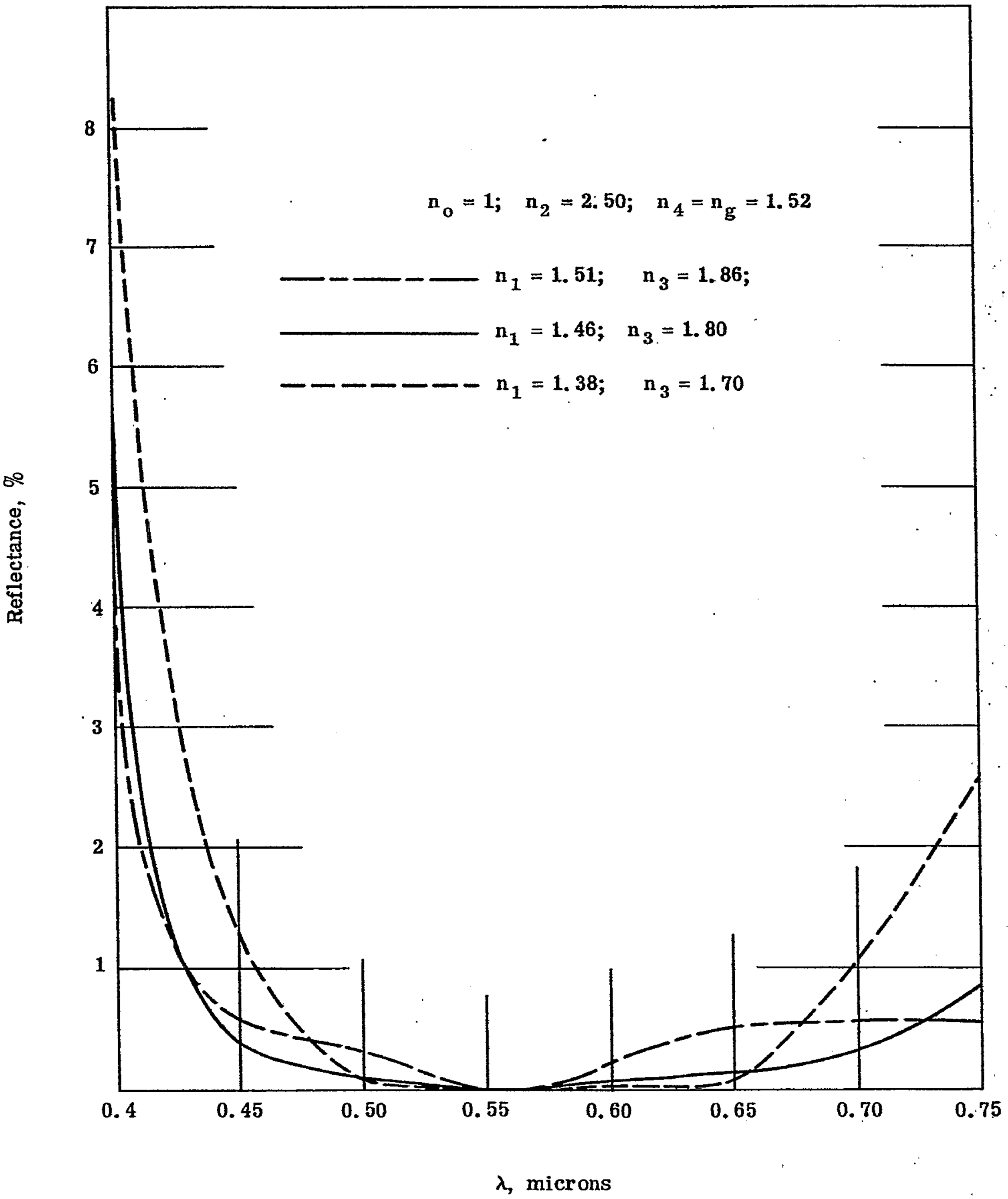


Figure 21.30- Spectral reflectances of quarter-half-quarter-wave trilayers on a substrate of refractive index 1.52. The high index $n_2 = 2.50$ belongs to the central, half-wave layer. The curves illustrate the effects of altering the refractive indices n_1 and n_3 of the quarter wave layers. The curves of this figure have been computed by the approximate method of section 21.2.11.

are properly identified with those of the trilayer. The condition for zero reflectance at $\lambda = \lambda_o$ is therefore

$$n_3 = \frac{n_1}{n_o}\sqrt{n_4}\,. \tag{199}$$

The refractive indices n_1 and n_3 of Figure 21.30 have been chosen in accordance with equation (199). The curves of Figure 21.30 are quite flat and low from 0.45 to 0.65 microns. They tend toward achromatization but do not succeed. The performances of the quarter-half-quarter and quarter-half-three quarter wave trilayers as anti-reflection films are not far different.

21.8.3 Band pass trilayers. A trilayer consisting of a dielectric layer silvered on both major surfaces is essentially a Fabry-Perot interferometer. Like Fabry-Perot interferometers, such trilayers are readily designed to transmit narrow bands of wavelengths after the manner described and illustrated in Section 16.16. These trilayers are utilized as narrow pass band filters. By forming the dielectric layer as a wedge of gradual taper, a convenient and effective monochromator is achieved. Thin film theory involves the assumption that the number of interreflections within each film is infinite. When regarded as Fabry-Perot interferometers, the trilayers discussed here require the choice of equation 16-(107) rather than 16-(106).

21.9 QUADRILAYERS

21.9.1 A low reflecting quadrilayer. The quarter-half-quarter trilayer discussed in Section 21.8.2 has been modified in a significant manner by Dr. Helen Jupnik so as to achieve a more practical anti-reflection film that has excellent performance. Difficulties occur in making the trilayers of Figure 21.30 because the refractive indices n_1 or n_3 or both are not available as durable, non-absorbing materials. To overcome difficulty due to availability of a material having the most desired refractive index n_3, Dr. Jupnik replaces the corresponding quarter wave layer by an "equivalent" quarter wave bilayer having refractive indices n'_3 and n'_4 as illustrated in Figure 21.31.

21.9.2 The principle of equivalence. The principle of equivalence is so important to the theory of thin films that its application to Jupnik's quadrilayer will be considered in detail. The substituted bilayer shall be a

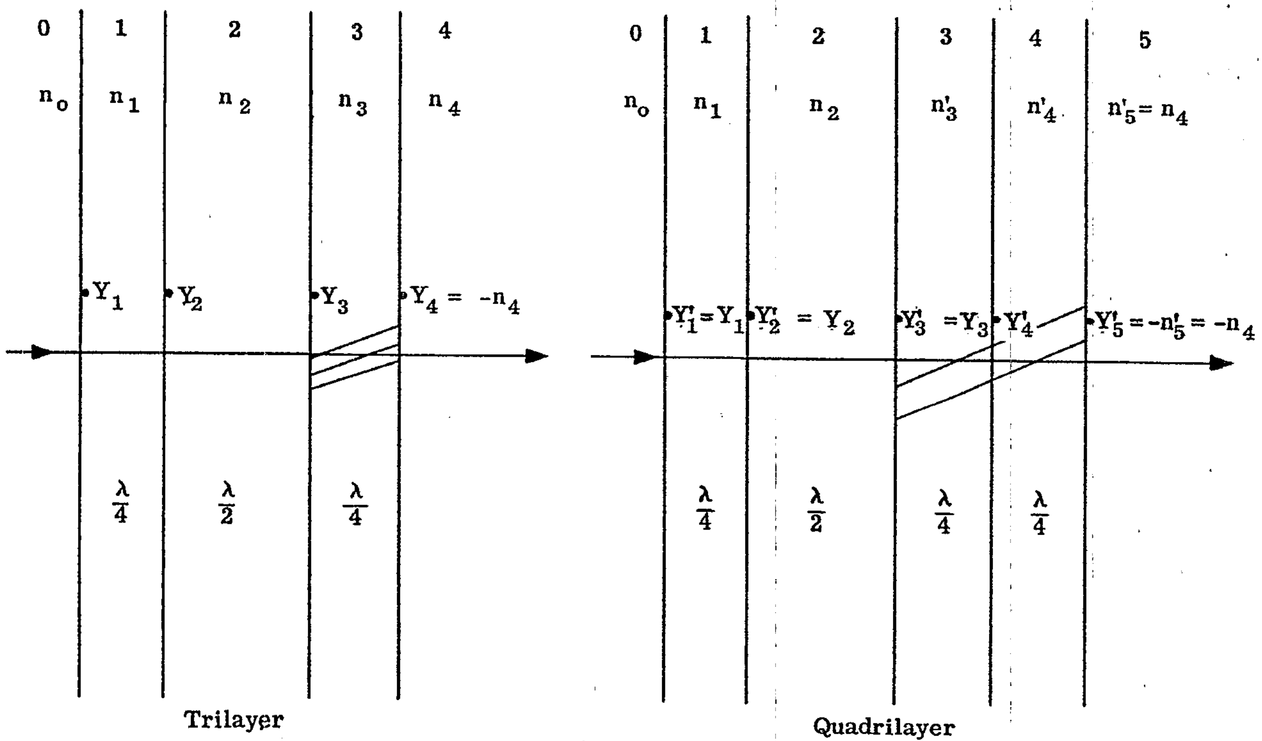


Figure 21.31- Notation with respect to the $\frac{\lambda}{4}-\frac{\lambda}{2}-\frac{\lambda}{4}$ trilayer and H. Jupnik's $\frac{\lambda}{4}-\frac{\lambda}{2}-\frac{\lambda}{4}-\frac{\lambda}{4}$ quadrilayer. The last quarter-wave layer is replaced by an equivalent bilayer. Y_ν denote admittances.

quarter wave bilayer and shall be equivalent to a quarter wave layer of refractive index n_3 at the wavelength λ_o at which the optical path is in fact one-quarter wavelength. The argument is simple on the basis of the admittances Y_ν. Suppose that it is possible to choose n'_3 and n'_4, Figure 21.31, such that $Y'_3 = Y_3$. Because layers number 1 and 2 have not been altered, it must now be true that $Y'_2 = Y_2$ and $Y'_1 = Y_1$, facts that can be checked, if desired, from equation (81a). From equation (79a) the complex reflectance ρ_o of the multilayer is $\rho_o = (n_o + Y_1)/(n_o - Y_1)$ at normal incidence. Hence ρ_o is left unaltered when $Y'_1 = Y_1$ or $Y'_3 = Y_3$. With respect to the trilayer $(N = 3)$, equation (81b) yields

$$Y_4 = -n_4. \qquad (200)$$

Since $\beta = \pi$ for quarter wave layers, equation (81a) yields

$$Y_3 = n_3^2/Y_4 = -n_3^2/n_4. \qquad (200a)$$

Similarly, with respect to the last two proposed elements of the quadrilayer, Figure 21.31,

$$Y'_5 = -n'_5 = -n_4; \qquad (200b)$$

$$Y'_4 = (n'_4)^2/Y'_5 = -(n'_4)^2/n_4; \qquad (200c)$$

$$Y'_3 = (n'_3)^2/Y'_4 = -n_4(n'_3)^2/(n'_4)^2. \qquad (200d)$$

By setting $Y_3 = Y'_3$ from equations (200a and d), one finds almost directly that

$$n'_3/n'_4 = n_3/n_4. \qquad (201)$$

The quarter wave bilayer having refractive indices n'_3 and n'_4 that satisfy equation (201) is equivalent to the quarter wave layer of refractive index n_3 for all values of n_3.

21.9.3 Selection of values.

21.9.3.1 We have seen that choosing n_3 in accordance with equation (199) makes $\rho_o = 0$ at $\lambda = \lambda_o$. By introducing n_3 from equation (199) into equation (201), we obtain Dr. Jupnik's selection for n'_3 and n'_4 in the form

$$\frac{n'_3}{n'_4} = \frac{n_1}{n_o\sqrt{n_4}}\ ; \quad n_4 = n'_5 = n_g. \qquad (202)$$

One is free to assign values to n'_3 or n'_4. In the example of Figure 21.32, n_1 and n'_4 are assigned the value 1.384 corresponding with the choice of $M_g F_2$. Then, with $n_o = 1$ and $n'_5 = 1.52$, one computes $n'_3 = 1.55$* from equation (202).

21.9.3.2 Comparison of Figures 21.30 and 21.32 reveals a number of interesting points. First, the substitution of the quarter wave bilayer serves also to achromatize the multilayer. Secondly, for the choice $n_1 = 1.38$ the spectral reflectance curve of Figure 21.32 is low and flat over a greater range of wavelengths. Thirdly, reflectances less than 0.1% are exhibited over a remarkably long range of wavelengths.

21.9.3.3 The effect of reducing the refractive index, n_2, of the half-wave member of the quadrilayer is illustrated in Figure 21.33. The achromatic points have moved outward to 0.45 and 0.70 microns to produce low reflectances over a greater portion of the visible spectrum than in Figure 21.32. This gain in spectral range is obtained at the cost of a slight increase of the reflectances at the points marked A and B. Further analysis shows also that the effects of increasing the refractive index, n_5, of the substrate are slight even when the refractive indices n_1 to n_4 of Figures 21.32 and 21.33 are left unchanged. From equation (202), a change in the refractive index of the substrate requires that n'_3 be changed correspondingly if one insists that the reflectance shall be zero at $\lambda = \lambda_o$.

21.10 QUARTER WAVE MULTILAYERS

21.10.1 Introduction. The following discussion is restricted to normal incidence upon non-absorbing systems and to the wavelength λ_o at which the optical path of each layer is one quarter wavelength. Equation (81a) shows that at $\lambda = \lambda_o$, where $\beta_\nu = \beta = \pi$,

$$Y_{\nu-1} = n_{\nu-1}^2/Y_\nu. \qquad (203)$$

---

* The refractive index of thorium oxyfluoride falls near 1.55.

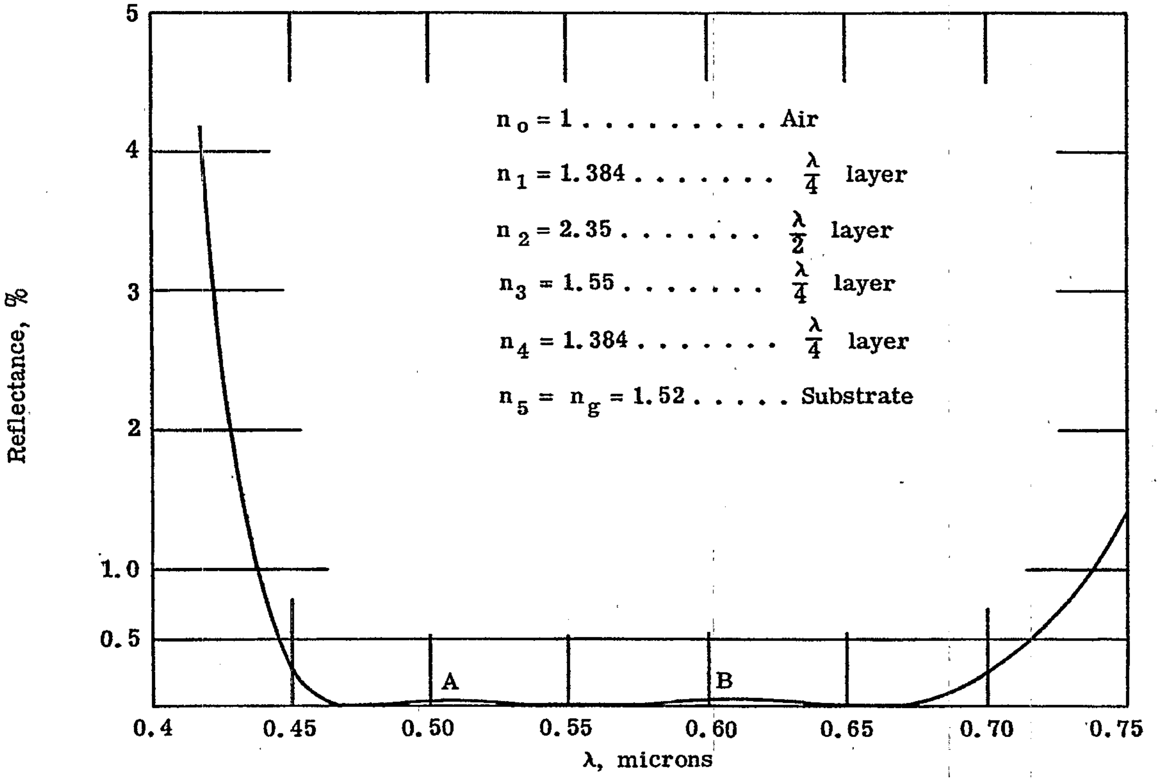


Figure 21.32- Curve of the computed spectral reflectances of a quadrilayer by Dr. H. Jupnik for the indicated refractive indices of the system.

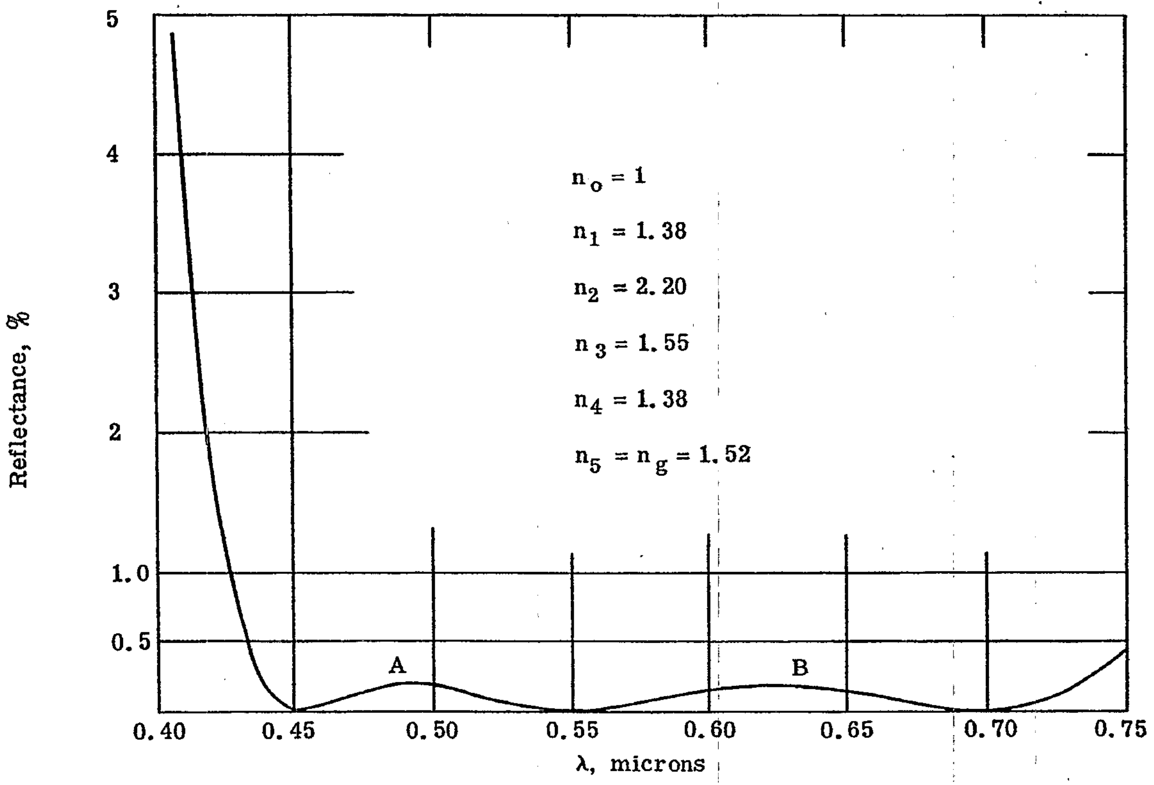


Figure 21.33- Curve of spectral reflectances illustrating the effect of decreasing the refractive index n_2 of the half-wave layer of the quadrilayer of Figure 21.32. The curves of both Figures 21.32 and 21.33 have been computed by an accurate method.

Closer examination of equation (81a) reveals that equation (203) holds whenever $\beta_\nu/2 = \beta/2 = \mu\pi/2$ where μ is an odd integer and $\beta/2$ is optical path. Increasing the optical path of any one or any number of the layers by an integral number of half-wavelengths will not alter the following conclusions.

21.10.2 The admittance, Y_1. From equation (203) one finds directly that $Y_2 = n_1^2/Y_1$; $Y_3 = n_2^2/Y_2 = Y_1 n_2^2/n_1^2$; $Y_4 = n_3^2/Y_3 = n_3^2 n_1^2/n_2^2 Y_1$; etc. Hence one concludes by induction that

$$Y_N = \frac{1}{Y_1} \frac{n_{N-1}^2 n_{N-3}^2 n_{N-5}^2 \cdot \cdots\cdots n_1^2}{n_{N-2}^2 n_{N-4}^2 n_{N-6}^2 \cdot \cdots\cdots n_2^2} \quad ; \; N \text{ even} ; \tag{204}$$

$$Y_N = Y_1 \frac{n_{N-1}^2 n_{N-3}^2 \cdot \cdots\cdots n_2^2}{n_{N-2}^2 n_{N-4}^2 \cdot \cdots\cdots n_1^2} \quad ; \; N \text{ odd} ; \tag{204a}$$

where N is the number of layers in the multilayer. However,

$$Y_N = n_N^2/Y_{N+1} = -n_N^2/n_{N+1} \tag{204b}$$

since $Y_{N+1} = -n_{N+1}$, see equation (81b). Hence, the admittance Y_1 of any quarter-wave multilayer is given by

$$Y_1 = -n_{N+1} \frac{n_{N-1}^2 n_{N-3}^2 \cdot \cdots\cdots n_1^2}{n_N^2 \; n_{N-2}^2 \cdot \cdots\cdots n_2^2} \quad ; \; N \text{ even} ; \tag{205}$$

$$Y_1 = -\frac{1}{n_{N+1}} \frac{n_N^2 \; n_{N-2}^2 \cdot \cdots\cdots n_1^2}{n_{N-1}^2 n_{N-3}^2 \cdot \cdots\cdots n_2^2} \quad ; \; N \text{ odd} . \tag{205a}$$

With Y_1 thus determined, the corresponding complex reflectance ρ_o of the multilayer can be computed from equation (79a). For normal incidence,

$$\rho_o = \frac{n_o + Y_1}{n_o - Y_1} . \tag{206}$$

21.10.3 The zero condition. According to equation (206), the reflectance will be zero at $\lambda = \lambda_o$, whenever $Y_1 = -n_o$, i.e. whenever

$$\frac{n_o}{n_{N+1}} = \frac{n_{N-1}^2 n_{N-3}^2 \cdot \cdots\cdots n_1^2}{n_N^2 \; n_{N-2}^2 \cdot \cdots\cdots n_2^2} \; ; \; N \text{ even} ; \tag{207}$$

$$n_o n_{N+1} = \frac{n_N^2 \; n_{N-2}^2 \cdot \cdots\cdots n_1^2}{n_{N-1}^2 n_{N-3}^2 \cdot \cdots\cdots n_2^2} \; ; \; N \text{ odd} . \tag{207a}$$

When $N = 2$, one obtains $n_o/n_3 = n_1^2/n_2^2$, the result of equation (194) for quarter wave bilayers. When $N = 3$, one obtains $n_o n_4 = n_1^2 n_3^2/n_2^2$, the zero condition for trilayers. Equations (207 and 207a) give one much greater flexibility as regards the choice of materials for obtaining zero reflectance than does the highly specialized zero condition for a monolayer or for a bilayer.

21.10.4 High reflecting multilayers. Equation (206) shows that there are two different ways in which one can achieve the result $|\rho_o| \to 1$. Thus the energy reflectance approaches its highest value unity as

$$Y_1 \to 0 ; \tag{208}$$

$$|Y_1| \to \infty . \tag{208a}$$

Suppose with respect to equation (205) that $n_{N-1}/n_N < 1$, $n_{N-3}/n_{N-2} < 1$, ... $n_1/n_2 < 1$. Then $|\rho_o| \to 1$ as $N \to \infty$ on account of equation (208). On the other hand, if one chooses $n_{N-1}/n_N > 1$, ... $n_1/n_2 > 1$, then $Y_1 \to \infty$ and $|\rho_o| \to 1$ as the number N of the layers in the multilayer approaches infinity. Similar observations apply when N is odd as in equation (205a). Therefore many possibilities are open for achieving high energy reflectance by increasing the number of layers in the multilayer.

21.10.5 The periodic system of repeated bilayers. The production of a multilayer is simplified by alternating layers of high and low refractive indices n_h and n_1. When the number N of layers is even, the resulting system of layers is periodic and forms an assembly of repeated bilayers as illustrated in Figure 21.34. There

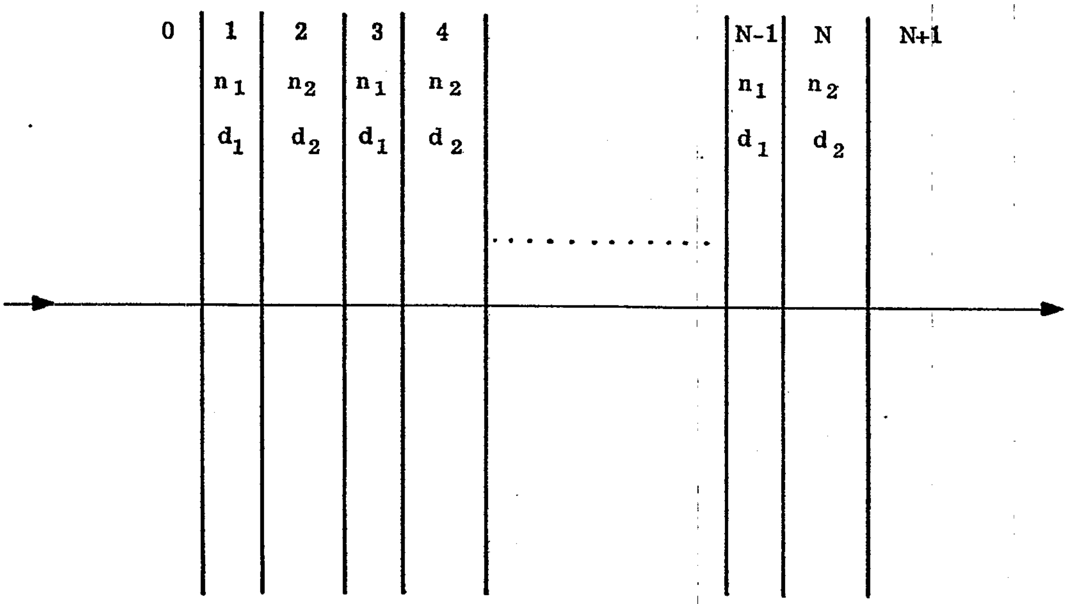


Figure 21.34- A multilayer consisting of repeated bilayers having the refractive indices n_1 and n_2. The thicknesses d_1 and d_2 are likewise repeated and are chosen so that the optical path is one quarter wavelength at an assigned wavelength λ_o. The number N of layers must now be even.

are N/2 even and N/2 odd integers between 0 and N when N is an even integer. Consequently, equation (205) simplifies to the result

$$Y_1 = - n_{N+1}\left(\frac{n_1}{n_2}\right)^N ; \tag{209}$$

and equation (206) assumes the explicit form

$$\rho_o = \frac{n_o - n_{N+1}\left(\frac{n_1}{n_2}\right)^N}{n_o + n_{N+1}\left(\frac{n_1}{n_2}\right)^N} = \frac{\frac{n_o}{n_{N+1}} - \left(\frac{n_1}{n_2}\right)^N}{\frac{n_o}{n_{N+1}} + \left(\frac{n_1}{n_2}\right)^N} . \tag{210}$$

The zero condition requires that

$$\frac{n_o}{n_{N+1}} = \left(\frac{n_1}{n_2}\right)^N . \tag{211}$$

If $n_o < n_{N+1}$, one must choose $n_1 < n_2$ in order to obtain zero reflectance, i.e. one must apply the layer of higher refractive index upon the substrate of refractive index n_{N+1}. On the other hand, equation (210) shows that $|\rho_o|$ can be made high whether one chooses $n_1 < n_2$ or $n_1 > n_2$ provided that N is taken sufficiently large; but that one should choose the alternative $n_1 > n_2$ in order to obtain the highest energy reflectance $|\rho_o|^2$ for a given number N of layers.

21.10.6 Odd number of alternating layers. - An important group of quarter wave multilayers containing an odd number N of layers having alternating refractive indices n_1 and n_2 is illustrated by Figure 21.35. If, for example, N = 5 and $n_1 > n_2$, these facts are indicated by the notation HLHLH or $(HL)^2$ H in which H and L refer to the high and low refractive indices n_1 and n_2, respectively. As a second example, if N = 15 and $n_1 < n_2$, the multilayer is described by writing $(LH)^7$ L. As in section 24.9.4, the complex

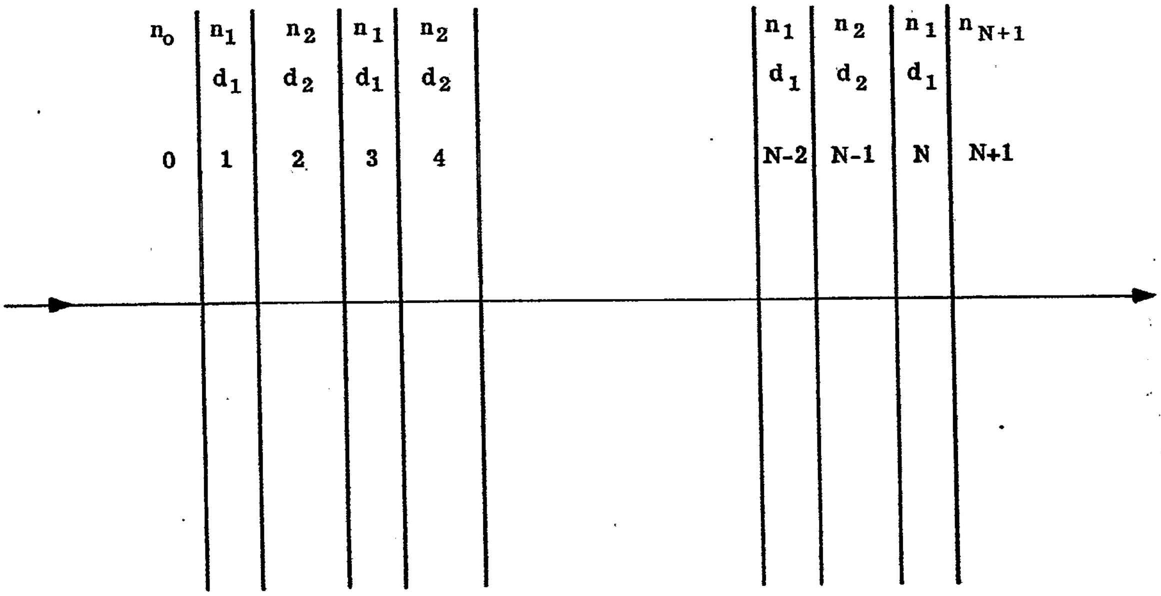


Figure 21.35- A multilayer consisting of an odd number N of quarter wave layers whose refractive indices n_1 and n_2 are alternated. A layer of refractive index n_1 occurs at each end of the multilayer. The optical paths of the layers are $\lambda/4$ at $\lambda = \lambda_o$.

reflectance ρ_o at $\lambda = \lambda_o$ can be found in a simple manner from equations (205a) and (206). The result is

$$\rho_o = \frac{n_o - \frac{n_1^2}{n_{N+1}}\left(\frac{n_1}{n_2}\right)^{N-1}}{n_o + \frac{n_1^2}{n_{N+1}}\left(\frac{n_1}{n_2}\right)^{N-1}}, \tag{212}$$

with the energy reflectance $R = |\rho_o|^2$. To obtain the highest value of R with the fewest number of layers, one chooses n_1 as large as possible and n_2 as low as possible. The curve of computed spectral transmittance $T = 1 - |\rho_o(\lambda)|^2$ of the multilayer $(HL)^5 H$ is shown in Figure 21.36 for the indicated values of n_o, n_1, n_2 and $n_{N+1} = n_g$. The layers are quarter wave layers at $\lambda = \lambda_o = 0.75$ microns at which one computes from equation (212) that $|\rho_o|^2 = 0.9946$ whence $T = 0.0054$. Energy reflectances exceeding those from silver are obtained easily with multilayers.

21.10.7 Achromatization. The spectral transmittance curve of Figure 21.36 illustrates one of the more serious difficulties encountered in the design of thin films. The oscillations in the spectral transmittances exemplified by those occurring between 0.4 and 0.62 microns are usually undesirable. The term achromatization or achromatizing pertains to the minimization of the amplitudes of undesired, rapid oscillations of spectral transmittance or reflectance curves in such a manner that the curves are flattened. This usage of the term achromatization is not entirely consistent with that of Section 21.7.4. Achromatization of bilayers and multilayers differ slightly, we may say, as to the manner in which a curve of spectral reflectance or transmittance is "flattened". An example * of one method of achromatizing multilayers is illustrated in considerable detail by Figures 21.35, 21.36, 21.37, and 21.38. Flattening of the curve of spectral transmittances between 0.4 and 0.62 microns is accomplished without appreciable alteration of the curve between 0.62 and 0.75 microns. As with bilayers, achromatization of multilayers can be achieved also by adding half-wave layers at strategic locations within the multilayer.

---

* It will not be possible due to lack of time and space to do justice to the able work of many investigators who have contributed to methods of achromatization.

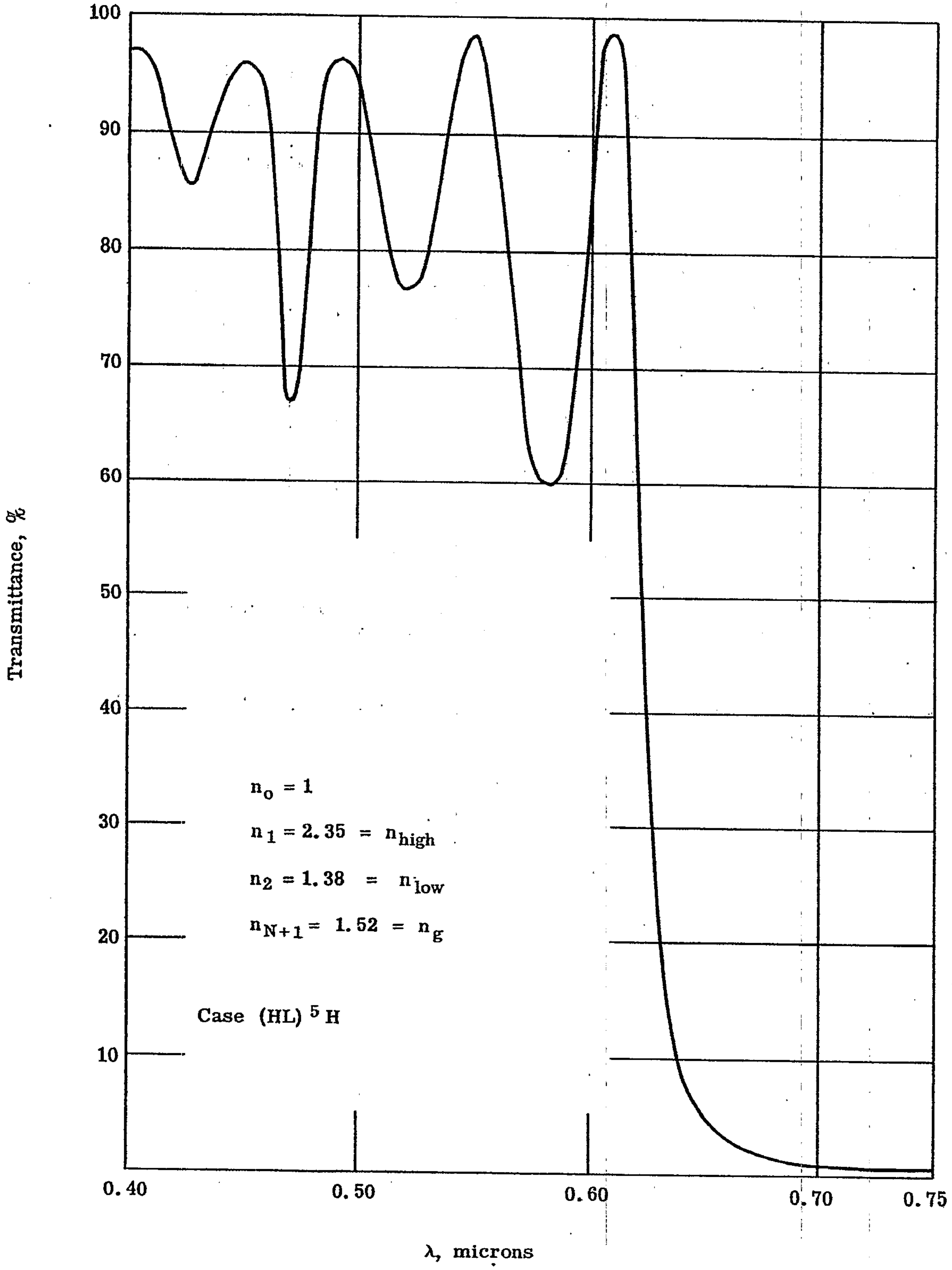


Figure 21.36- Curve of computed spectral transmittances taken from the files of Dr. H. Jupnik for the alternating multilayer $(HL)^5$ H having layers that are quarter-wave layers at $\lambda_o = 0.75$ microns. The heights and locations of the numerous maxima and minima have not been determined with greatest possible care.

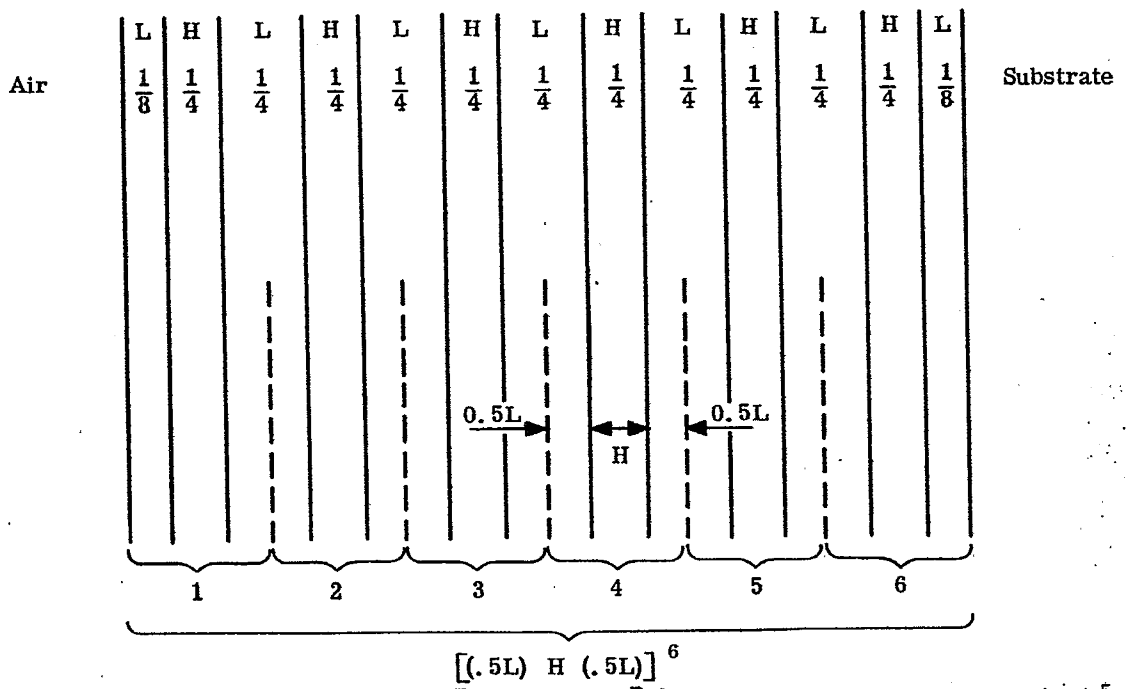


Figure 21.37- Explanation of the notation $[(.5L)\ H\ (.5L)]^6$ with respect to a multilayer $(HL)^5H$ to which has been added at each end an achromatizing layer of low refractive index and of optical path $\lambda/8$ at $\lambda = \lambda_0$. The system consists of six repeated "trilayers," $(.5L)\ H\ (.5L)$ or $(\frac{L}{2})\ H\ (\frac{L}{2})$.

21.10.8 Narrow pass band filters.

21.10.8.1 The design of multilayers that are intended for use as narrow pass band filters is based upon the principles of the Fabry-Perot interferometer. The silver coatings on opposite surfaces of the plane parallel plate are replaced by high reflecting multilayers. In turn, the plane parallel plate is usually replaced by a layer that serves as the spacer. High reflecting multilayers are superior to silver coatings * when only small amounts of absorption can be tolerated and when durability becomes an important consideration. One arrangement is illustrated in Figure 21.39. When all the layers of multilayers, B_1 and B_2, are quarter wave layers at $\lambda = \lambda_0$, the optical path of the separating layer, S, should be an integral number, ν, of half waves at λ_0. It is not difficult to see that at normal incidence the transmittance is unity at $\lambda = \lambda_0$ for the idealized system that contains no absorption, scattering or departures from the rigid design of Figure 21.39. First, one notes that the spacer layer S is an absentee layer at λ_0. Layers 1 and 1' are then effectively in contact and comprise a half-wave or absentee layer. This now places layers 2 and 2' effectively in contact to form a third absentee layer. One concludes that all opposing pairs of layers form absentee layers at λ_0 -- and, indeed, that the filter becomes an "absentee filter" that must have transmittance unity.

21.10.8.2 The behavior of the filter at $\lambda \neq \lambda_0$ can be appreciated and evaluated as follows from the theory of Fabry-Perot interferometers. With reference to equation 16-(103) we note that the parameter, A, is now given by

$$A = |\rho_0| \quad |\rho'_0| \tag{213}$$

in which ρ_0 and ρ'_0 are the complex reflectances of the "coated" surfaces of the spacer, S, Figure 21.39, since the spacer has been assumed to be non-absorbing. Let T denote the time averaged energy transmittance of the filter. Then T = W where W is the quantity given by equation 16-(107). From this equation we see that

$$2T = \frac{|\tau \ \tau'|^2}{1 - 2A\cos\alpha + A^2} \tag{214}$$

---

* Silver coatings having extremely small amounts of absorption can be produced by evaporation; but the required technique is not well known and the silver coatings are not likely to remain low-absorbing.

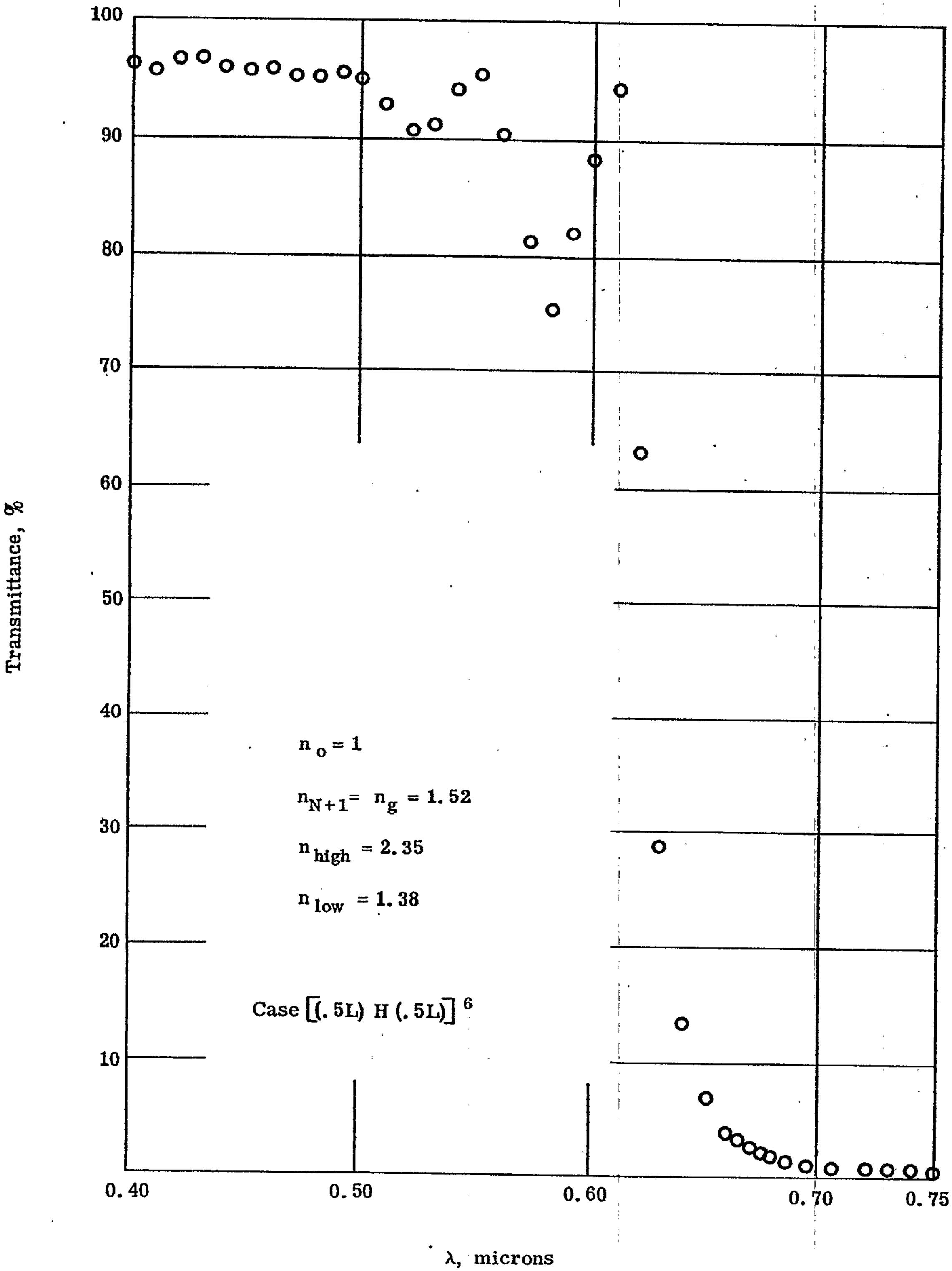


Figure 21.38- Plot of the spectral transmittances obtained by achromatizing the system $(HL)^5H$ of Figure 21.35. by the addition of a λ/8 layer of low refractive index at each end of the multilayer. The notation for the multilayer thus obtained is $[(.5L)\ H\ (.5L)]^6$. These plotted data have been taken from the files of Dr. H. Jupnik.

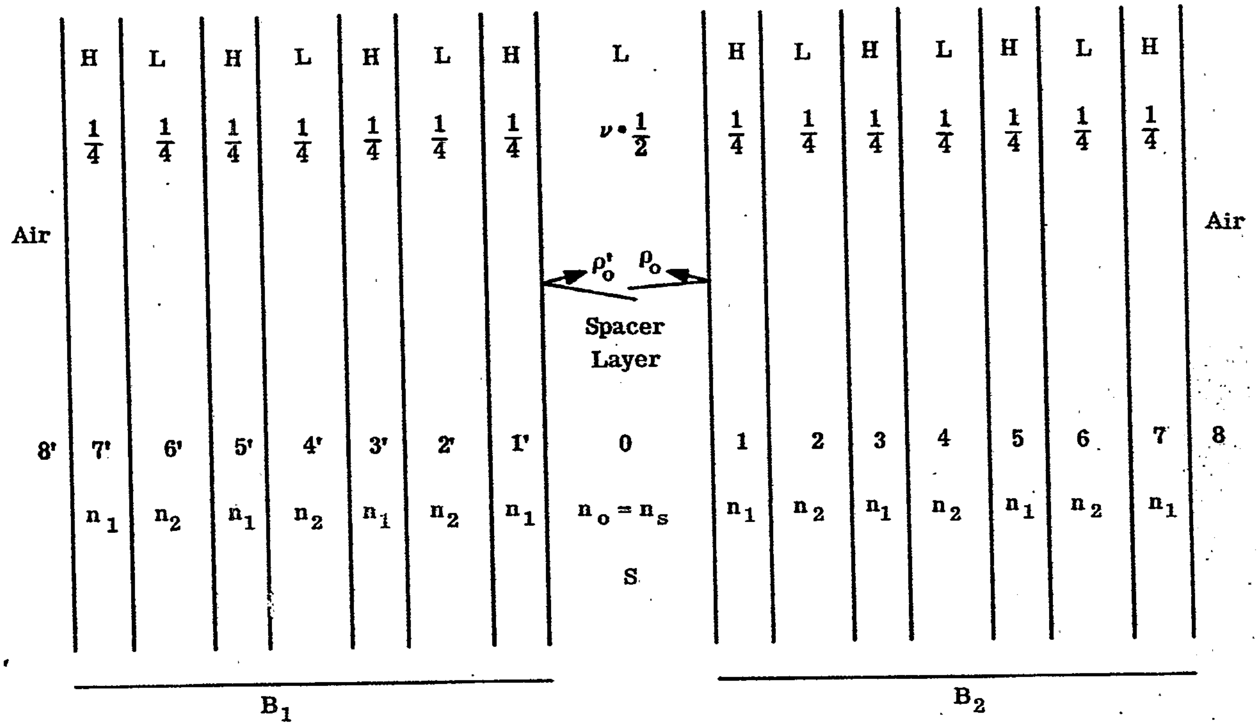


Figure 21.39 - A narrow pass band, interference filter consisting of a spacer layer coated with high reflecting, quarter-wave multilayers B_1 and B_2 that consist of seven alternating layers.

in which τ' and τ are the complex transmittances of multilayers B_1 and B_2, Figure 21.39, and

$$\alpha = \frac{4\pi n_s d_s}{\lambda} + \arg(\rho_o) + \arg(\rho'_o) \tag{215}$$

where n_s and d_s are the refractive index and thickness, respectively, of the spacer. The transmittance T is maximum when

$$\alpha = \nu 2\pi \tag{216}$$

where the integer ν is called the order number of the filter or interferometer. The corresponding wavelength is usually indicated by λ_ν in Fabry-Perot interferometry. By eliminating α from equation (215) with the aid of equation (216), one obtains the result

$$2n_s d_s + \frac{\lambda_\nu}{2\pi}\left[\arg(\rho_o) + \arg(\rho'_o)\right] = \nu\lambda_\nu . \tag{217}$$

Sharp peaks of transmittance $T = W$ are produced by equation (214), as illustrated by Figure 16.19, when the equivalent equations (216) and (217) are satisfied. These sharp peaks are the narrow pass bands.

21.10.8.3 Suppose now that the numerous conditions of Figure 21.39 are met by the interference filter at $\lambda = \lambda_o$. Then $\rho_o = \rho'_o$ and ρ_o is determined from equation (212) with $N = 7$. Because we have chosen $n_1 > n_2$, ρ_o turns out to be real and negative. We may write

$$\rho_o = |\rho_o| \, e^{\pm i\pi} . \tag{218}$$

The phase change on reflection can be taken as either of the two physically indistinguishable values, $\pm\pi$. The simplest way of interpreting equation (217) at $\lambda = \lambda_o$ is now to take $\rho_o = |\rho_o| \, e^{i\pi}$ and $\rho'_o = |\rho_o| \, e^{-i\pi}$ since these are physically indistinguishable. Correspondingly,

$$\arg(\rho_o) + \arg(\rho'_o) = 0 ; \quad \lambda = \lambda_o . \tag{219}$$

Then, simply,

$$2n_s d_s = \nu \lambda_\nu \quad \text{at} \quad \lambda = \lambda_o . \qquad (220)$$

Hence we may regard the integers ν of Figure 21.39 and equation (220) as the same integer, i.e. as the spectral order number of the filter. Equation (220) explains why an interference filter can exhibit two pass bands in the visible region. Suppose that λ_o is chosen at the red end of the spectrum. Here, $\lambda_o = \lambda_\nu$. If ν is high enough, it can happen that at a shorter wavelength $\lambda_{\nu+1}$ the order number is increased to $\nu + 1$. In other words, it can happen that

$$2n_s d_s = (\nu + 1)\lambda_{\nu+1} , \qquad (221)$$

where $\lambda_{\nu+1} < \lambda_o$. Because equation (217) and its simplifications are conditions for maxima of T, the wavelength $\lambda_{\nu+1}$ defines the center of a subsidiary pass band. By increasing ν (increasing d_s), one obtains additional pass bands in a specified spectral range such as the visible region.

21.10.8.4 The widths of the narrow pass bands are important properties of the interference filter. These widths can be evaluated as follows. From equation 16-(113),

$$|\Delta \alpha| = \frac{1 - A}{\sqrt{A}} , \qquad (222)$$

where $\Delta\alpha$ is a particular departure of α from its value α_ν at the wavelength λ_ν at which T is maximum. This departure of α from α_ν changes T from $T = T_{maximum}$ to $T_{maximum}/2$. Let us assume for simplicity that the variation of the phase changes on reflection $\arg(\rho_o)$ and $\arg(\rho'_o)$ with wavelength can be ignored.* By differentiating α with respect to λ in equation (215), one obtains

$$|\Delta\alpha| = \frac{4\pi n_s d_s}{\lambda^2} |\Delta\lambda| . \qquad (223)$$

We evaluate the derivative at the center of the pass band under consideration where $\lambda = \lambda_\nu$. Let $2n_s d_s$ be eliminated from equation (223) with the aid of equation (220). Then

$$|\Delta\alpha| = \nu 2\pi \frac{|\Delta\lambda|}{\lambda_\nu} . \qquad (224)$$

By eliminating $\Delta\alpha$ from equation (224) with the aid of equation (222), we obtain the formulae

$$2|\Delta\lambda| = \frac{\lambda_\nu}{\nu} \frac{1 - A}{\pi\sqrt{A}} ; \qquad (225)$$

$$= \frac{\lambda_\nu}{\nu} \cdot \frac{1 - |\rho_o| \; |\rho'_o|}{\pi\sqrt{|\rho_o| \; |\rho'_o|}} ; \qquad (225a)$$

$$= \frac{\lambda_\nu}{\nu f} \qquad (225b)$$

where the finesse (16) f is defined as

$$f = \pi\sqrt{A}/(1 - A) . \qquad (226)$$

$|\Delta\lambda|$ is the half-width of the pass band of the filter at the selected order number ν. By definition, $|\Delta\lambda|$ is the departure of λ from λ_ν that causes the transmittance of the filter to drop from $T_{max.}$ at λ_ν to $T_{max.}/2$ at $\lambda = \lambda_\nu \pm |\Delta\lambda|$. When $|\rho_o| = |\rho'_o|$,

$$2|\Delta\lambda| = \frac{\lambda_\nu}{\nu} \frac{1 - |\rho_o|^2}{\pi|\rho_o|} , \qquad (227)$$

where $|\rho_o|$ is evaluated at the wavelength λ_ν. At the wavelength $\lambda = \lambda_o$, ρ_o can be calculated from equation (212) provided that the high reflecting multilayer falls in the class governed by equation (212). Equation (227) shows that the half-width is decreased by increasing $|\rho_o|$ or the order number ν. At fixed λ_ν, the order number is increased by increasing the optical path $n_s d_s$ of the spacer.

---

* In some applications, $\arg(\rho_o)$ is deliberately made a rapid function of wavelength in order to utilize the dispersion of the phase changes on reflection for obtaining narrow pass bands.

(16) P. Giacomo, Rev. D'optique, 35, 317 (1956).

21.10.8.5 As an example of the evaluation of a half-width from the theory, let us now consider the type of interference filter of Figure 21.39 with $n_1 = 2.35$, $n_2 = 1.38$ and $n_s = 1.38$. This choice corresponds to the use of ZnS and MgF_2 as the material of high and low refractive index, respectively. We suppose that the multilayers B_1 and B_2 are quarter-wave systems at $\lambda_o = 5550$ Angstroms and that the spacer is a half-wave spacer at λ_o. We compute the half-width at the main transmittance peak located at $\lambda_\nu = \lambda_o$ where $\nu = 1$. In applying equation (212), we may take $n_o = n_2 = 1.38$, $n_{N+1} = 1$, $n_1 = 2.35$ and $N = 7$. Then $|\rho_o| = 0.9797$ and $|\rho_o|^2 = 0.9598$. Then from equation (227), $|\Delta\lambda| = 36.2$ Angstroms. The half-width may be decreased to 12.07 Angstroms by choosing $\nu = 3$. In practice, further reduction of $|\Delta\lambda|$ is accomplished by increasing $|\rho_o|$, i.e. by increasing N or by choosing a pair of dielectric materials for which the ratio n_1/n_2 is higher. We see that half-widths of one Angstrom or less become difficult to attain.

21.11 MATERIALS AND TEXTS

Whereas many scattered publications deal with the optical properties of substances that are suitable for use as thin films, the writer is unaware of a publication that contains an exhaustive summary of the optical properties of the many possible materials from the ultraviolet region into the region of the infrared.

One of the longest tables of the optical and mechanical properties of materials that are used in making thin films will be found in L. Holland, "Vacuum Deposition of Thin Films," John Wiley & Sons, Inc. (1956).

Quite detailed discussion of the optical constants and properties of metallic films is included in O. S. Heavens, "The Optical Properties of Thin Films," Butterworths Scientific Publications, London (1955). This book includes scattered information about the optical constants of other materials such as ZnS, Sb_2O_3, Ge and Te.

A useful list of the optical constants of metals and inorganic compounds appears in most editions of "Handbook of Chemistry and Physics," Chemical Rubber Publishing Co.

A book by W. Lewis, "Thin Films and Surfaces," Temple Press Ltd, London, First Edition is devoted to the structure, properties and production of various thin films. Emphasis is placed upon aluminum and alloys containing aluminum.

The excellent work of Dr. Georg Hass and his associates has contributed information about the optical properties and formation of thin films -- especially the oxides of titanium, silicon, aluminum and rare earths. As one example, see Georg Hass, Vacuum, 2, 331-345 (1952). This publication contains a substantial list of references.

The following texts may be consulted for much additional, valuable discussion relative to thin films.

Auwarter, Max, ed. -- "Ergebnisse der Hochvakuum technik und der Physik dunner Schichten," Stuttgart, Wissen schaftliche Verlagsgessellschaft (1957).

Mayer, Herbert -- "Physik dünner Schichten," Stuttgart, Wissenschaftliche Verlagsgesellschaft, 1950, Volume 1 and 2.

Vasicek, Antonin -- "Optics of thin films," Amsterdam, North-Holland Publishing Co., 1959.

22 INFRARED OPTICAL DESIGN

22. 1 INTRODUCTION

22. 1. 1 General. The basic principles of optical design for the infrared region are the same as those for visible and ultraviolet light. The differences arise mainly from the nature of the materials which must be used, and from the operational and environmental requirements of most of the current applications.

22. 2 INFRARED OPTICAL MATERIAL

22. 2. 1 Image converter tube. Reflecting and refracting materials suitable for use at the various infrared wavelengths have been discussed in Section 16. In particular, reference was made to the publication by Ballard, McCarthy, and Wolfe, tabulating information on currently available materials. (Development work is active in this field, and the designer should keep abreast of the situation with appropriate journals and other sources of possible information on new materials). Only general comments on materials will be made here, although they will be extended somewhat in the subsequent portions of this section. Radiation in the 0. 8 to 1. 2μ region is used with night vision devices employing image converter tubes, such as the "Sniperscope". These are ordinarily "active" devices. That is, they are used to look at objects which are illuminated by infrared light from a source which is under control of the user. The light source is usually a tungsten lamp or carbon arc covered by a filter which absorbs the visible light while passing the infrared. The effective wavelength range results from the combination of the spectral characteristics of the source, of the filter, and of the photo-sensitive cathode of the image converter tube.

22. 2. 2 Infrared imagery. In infrared use, an objective similar to a photographic objective forms an image, in infrared light, on the photocathode of the tube. The illuminated areas of the cathode emit electrons which are accelerated and focussed on a fluorescent screen at the opposite end of the tube, thus forming a visible image which can be viewed by the user with the aid of a magnifier.

22. 2. 3 Glass for infrared usage. Ordinary optical glass transmits satisfactorily in this region and is always used. However, the dispersion characteristics of the several types become much more nearly alike in the infrared than they are in the visible and, as a consequence, much stronger powers of crowns and flints are required to obtain achromatization. For example, a doublet with a 100mm. focal length, made of light barium crown (1. 5725/57. 4) and dense flint (1. 6170/36. 6), and achromatized in the visible, will have a crown with a 36mm. focal length and a flint with a focal length of 57mm. If the same glasses be used to achromatize the doublet in the region from 0. 8μ to 1. 2μ, the crown must have a focal length of 19mm. and the flint 23mm. The chromatic aberration of a single crown lens in this region is approximately two-thirds to three-quarters of that for a lens of the same glass from the C line to the F line in the visible, and some slight advantage can be taken of this fact. It is still true, however, that in using a basic lens type, e. g., a Petzval, it is sometimes necessary to replace a single crown lens by two, in order to avoid the high-order aberrations which would otherwise result from the strong curves necessary for achromatization.

22. 2. 4 Optical glass infrared absorption. Radiation in the region from the visible to about 3. 5μ is within the range of usefulness of lead sulfide cells. Ordinary optical glass begins to absorb slightly at about 2. 0μ, and the absorption becomes very great at approximately 2. 6μ to 2. 7μ, depending on the type. Consequently, ordinary glass cannot be used for systems requiring performance beyond 2. 7μ. Although the usefulness of lead sulfide cells extends to 3. 5μ or beyond, it may happen that the combination of (1) the spectral characteristics of the source, (2) any filters in the system, (3) the intervening medium such as air, and (4) the lead sulfide cell itself, will produce a situation under which only a very small proportion of the response of the cell would be lost by using ordinary glass. In such a case, it is worth while to consider carefully whether the loss of a slight amount of response is sufficient to outweigh the advantages of using ordinary glass. It is well to study a number of glasses in this connection, since there is some variability in transmission from type to type, remembering that the flints as a class, transmit slightly better than the crowns.

22. 2. 5 Materials suitable for wavelengths beyond 2. 7μ. The materials available and suitable for use at wavelengths beyond approximately 2. 7μ have, for the most part, refractive characteristics quite different from those with which the designer works in the visible and the near infrared. Indices of refraction range from approximately 1. 35 (lithium fluoride) to 4. 1 (germanium). The range of dispersion characteristics is even more striking. With ordinary glass, the ration of ν values available for achromatization is limited to about 2. 4:1 or less. In the infrared, this range may run as high as 46:1, the value for a positive silicon element and a negative element in the 3. 5 - 5. 5μ region. In spite of this great range of values of optical constants, it usually turns out that for reasons not primarily optical, only a few media are available for a given application. The designer must make his selection from those few, and determine

by trial which combination allows him to get the best correction within the limits of allowable complexity of the system. For this reason, among others, reflecting systems are frequently used. The reflectors are completely achromatized, of course, and have nearly constant reflectance over a large range of wavelength. Refracting elements are used with the reflectors to control aberrations.

22. 3 ENVIRONMENTAL REQUIREMENTS

22. 3. 1 Current applications. Currently, most designs of infrared optical systems are intended for use by the Military, and therefore the systems must meet Military requirements for serviceability after long exposure to adverse environmental conditions. In particular, many infrared systems are to be airborne, and must meet stringent requirements for compactness and lightness of weight, as well as the ability to withstand rapid changes in temperature and humidity without damage. These requirements place more rigid limitations on the choice of materials, and on the elaborateness of system, than are encountered in other applications.

22. 3. 2 Choice of materials. The choice of material for the windows of airborne equipment is one which must be based primarily on considerations of this nature. These windows are sometimes flat, but are more frequently dome-shaped, since they are used with scanning equipment. Being exposed on the exterior of the vehicle, they must be resistant to the variations of temperature and humidity to be expected in service. The material must be hard enough to withstand excessive scratching; for example, from dust kicked up during take-off and landing. Particularly when used in supersonic vehicles, the material must be able to withstand the thermal shock resulting from friction with the atmosphere. When heated from such friction, it must not radiate much energy in the infrared region in which the optical system is operating; otherwise a false signal may be generated or a true one obscured. Of course, a material radiates only in the region in which it absorbs (See Section 16) and therefore, ordinarily a material transmitting well enough to be considered for use as a window will not give trouble in this respect. However, because of the scarcity of suitable materials, it is sometimes necessary to consider those which do have slight absorption in the region of use and the possible effect of such unwanted radiation must be considered. In this connection, it is important to know the transmission characteristics of the material at elevated temperatures, since these characteristics may differ significantly from those at ordinary temperatures.

22. 3. 3 Size limitations. The window material must be obtainable in pieces large enough for the intended use. This requirement frequently rules out a number of otherwise promising materials. Some attempts at getting around this difficulty have been made by using segmented windows made up of a number of small pieces, and by replacing domes by polyhedrons made up of small, flat pieces. Such structures do not seem to have been generally satisfactory, however, probably because of the difficulty of providing adequate strength in combination with freedom from excessive obscuring by the supporting framework.

22. 4 OPERATIONAL REQUIREMENTS

22. 4. 1 Detection of infrared. To be useful, the infrared optical system must feed the energy it collects into a photosensitive device of some type. Fundamentally then, the design of the complete infrared instrument requires similtaneous consideration for the optics, the photosensitive device, and the associated equipment (usually electronic in nature), with respect to the performance requirements to be met. With respect to this discussion, photosensitive devices will only be described for the purposes of orientation.

22. 4. 2 Classification by instruments. Infrared devices are customarily classified by systems engineers as "image forming", or "non-image forming". An "image forming" device is an instrument whose output is a visual pictorial display of the field viewed by the device. An example is the "Sniperscope" previously mentioned. A "non-image forming" device is an instrument whose output is a signal, which is usually electrical. An example is an instrument giving information of the presence of a target of some nature in a particular portion of the field of view. This classification, while logical from the systems engineer's point of view, is not always very significant to the optical designer, since the optical systems of many "non-image forming" devices must actually have an optical image somewhere in the system in order to permit the location of a target within a particular portion of the field of view. Similarly, some "image-forming" devices, which depend on scanning procedures, require optical systems which simply condense the energy from a small field onto a photocell.

22. 4. 3 Classification by wavelength range. For the optical designer, a more useful classification of infrared devices is by the wavelength range which the instrument utilizes. For the purpose of this discussion, the range of the near infrared will include the region from 0.75μ to 3.0μ and beyond the near infrared will include the region from 3.0μ to 1000μ.

22.5 THE NEAR INFRARED REGION

22.5.1 Current applications. There are three types of commonly used systems which work in the near infrared; that is, in the region from the visible to about 1.3μ. All three are image formers, both in the systems sense and in the optical sense. They are the infrared photographic process, the image-converter tube systems, and the triggered radiation system. Ordinary optical glass is suitable at these wavelengths and the optical design is quite similar to that for photographic objectives, within limitations discussed in 17.7.2

22.5.2 Infrared photography. Infrared photography uses plates or films similar to those of photography with visible light, except that the emulsions have been sensitized by the addition of infrared-sensitive dyes.

22.5.3 Infrared image converter systems.

22.5.3.1 The system using an infrared image converter tube has an optical objective which forms an infrared image on the photosensitive cathode of the image converter as shown in Figure 22.1. The cathode emits electrons into the space within the tube, the rate of emission from a given area being proportional to the intensity of illumination of the area. The electrons are focused by electrostatic or electromagnetic means, in order to form an image on an electron-sensitive phosphor at the other end of the tube. The phosphor emits visible light, and the image can be seen by viewing the phosphor with the eye, usually with the aid of a magnifier.

22.5.3.2 Specifications for the components of the system are determined by a compromise between the desired performance characteristics and the state of the tube-maker's and the optical designer's arts. There are usually rather stringent requirements for compactness and portability. The instrument must operate as a telescope of a certain power, usually unity or greater. It is desirable to have a fast objective, since this increases the range at which the instrument is effective. Given the desired field angle, the size of the cathode of the tube determines the focal length of the objective, or vice versa. The electrostatic tubes (which are the sort usually used in this country) operate at a magnification less than unity. The sub-system consisting of objective and tube can then be considered as having an equivalent focal length equal to the focal length of the objective multiplied by the magnification of the image tube. (Both the objective and the electrostatic tube invert the image, so an erect image is presented on the phosphor.) For example, if the system is to have 1-1/2X power, the objective has a focal length of 50mm, and the image converter has a magnification of 0.7X; then the focal length

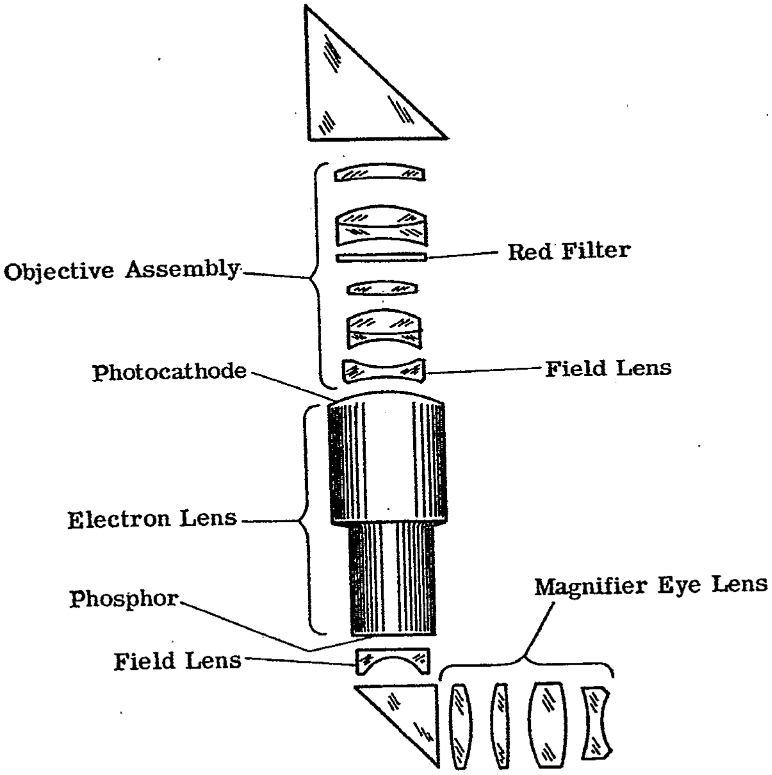


Figure 22.1- Optical schematic of an image converter system.

of the front system is 0.7X . 50mm = 35mm, and the magnifier must have a focal length of about 23mm.

22.5.3.3 The photocathode usually must be rather sharply convex toward the incident light, since the electronic, as well as the optical system, has field curvature. Thus, its curvature is of the opposite sign to that necessary to match the natural curvature of field of a refracting objective. For this reason, a strong negative field lens is employed in front of the cathode. The required power is such that it may be difficult to obtain the required correction for curvature with a single lens without getting total internal reflection of the light from the outer portions of the field; the work must then be divided between two elements. In addition to correcting the curvature, the field lens introduces coma, astigmatism and distortion, since it is not located exactly at the image. The coma and astigmatism must be balanced in the main part of the objective.

22.5.3.4 The electronic system of the image tube produces strong pincushion distortion in the image on the phosphor. This distortion, together with that of the objective and field lens, is dealt with, if at all, in the magnifier through which the phosphor is viewed.

22.5.3.5 The viewing lens system must be considered as a magnifier rather than an eyepiece. (See Section 13) Since the phosphor is a self-luminous surface, it emits light in all directions, and there is no natural exit pupil such as is present in an ordinary viewing telescope. This is an advantage in that there is more freedom to position the eye of the observer than would be the case with the presence of an exit pupil. However, it results in the necessity of correcting the magnifier for spherical aberration and coma, in order to avoid weird distortions and blurrings which can occur if the observer's eye does not happen to be exactly on axis. (This correction can be much cruder than necessary in an objective, since the pupil of the eye accepts only a portion of the bundle from a given object point at any one instant).

22.5.3.6 The conditions of use are such that it may be advantageous to use one or more aspheric surfaces in the magnifier. Aspherics of sufficient precision can be made by processes suitable for mass production. Such a magnifier is shown schematically in Figure 22.2. The magnifier consists of an eye lens and a field lens. One surface of the eye lens is aspherized to correct for spherical aberration. The bending of the lens is chosen to minimize coma. The field lens is aspherized to compensate for the pincushion distortion at the phosphor. (The aspheric may be given a slight power on axis to facilitate fabrication.)

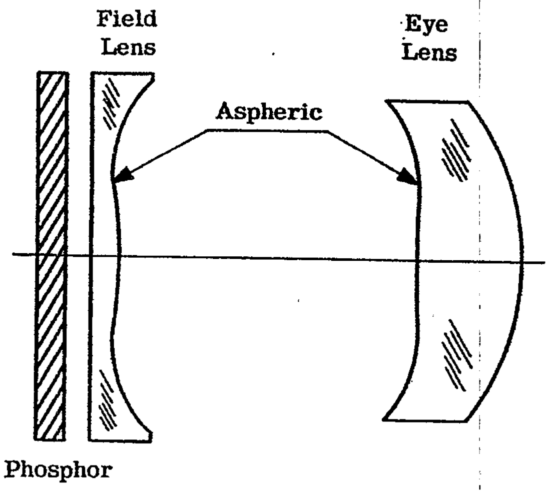


Figure 22.2- Aspheric Magnifier.

22.5.4 The triggered radiation type.

22.5.4.1 Instruments of the this type depend on the ability of some phosphors to store energy when irradiated by short wave radiation, and to emit it as visible light when triggered by irradiation with infrared. The short-wave radiation may be ultraviolet (or visible) light, or that from a bit of radioactive material. An objective forms the infrared image of the scene being viewed on the phosphor. The various parts of the phosphor emit visible light in proportion to the intensity of the infrared radiation.

22.5.4.2 In this system, it is necessary to provide optical means of inverting the image, since the system would otherwise perform like an inverting astronomical telescope. This may be done by using a lens or a prism erecting system, either preceding the phosphor (thus working in the infrared) or between the phosphor and the eye. (Other ingenious means have been used in special designs).

22.6 THE INTERMEDIATE AND FAR INFRARED REGION

22.6.1 Distinguishing characteristics.

22.6.1.1 Design in the region beyond the near infrared has two main distinguishing characteristics, in addition to the necessity of using materials other than optical glass. One is the limitation on image quality. The other is the limitation imposed by the combination of the performance requirements and the characteristics of the types of energy detectors which must be used in many applications.

22.6.1.2 Aside from these limitations, the design requirements are much as they are in other wavelength regions, and the designer must be prepared to deal with requirements quite similar to those found in other optical designs.

22.6.2 Limitation on image quality.

22.6.2.1 Since the diffraction limit of resolving power (see 16.28) depends on the wavelength of the light being used, the best image quality obtainable from a source with a given aperture is much poorer in the infrared than in the visible. Taking 0.56μ as typical of the visible region, the resolution at 3μ is five times as coarse, and at 10μ is eighteen times as coarse as in the visible. Since there is no gain to be achieved from improving the correction beyond the point at which the Rayleigh criterion is satisfied, the designer may stop his work with a residue of aberration which it might be well worth while to remove if he were working in the visible region. He may also have to warn the proposer of the system of the limited resolving power which can be obtained.

22.6.2.2 Currently the resolution requirements of most systems are even coarser than the limit which would be imposed by diffraction. However this is not always the case, even at present, and as the infrared art develops it is likely that there will be many more requirements for performance near the diffraction limit.

22.6.3 General functions of the optical system.

22.6.3.1 As an aid in discussing the relation of the energy detector to infrared optical design it is worth while to review the functions of the optical instrument in general terms.

22.6.3.2 Every optical instrument is designed to obtain information concerning the radiation characteristics of a portion of space. This portion of space is called the field of view of the instrument. It may be, for example, the crater of an arc, as in emission spectroscopy; a volume of space, as in absorption spectroscopy or in an infrared search system; or a surface, as in a slide projector. The radiation may or may not originate in the field; that is, the field may or may not be self-luminous. In emission spectroscopy and pyrometry the field is self-luminous. In absorption spectroscopy and with active viewing systems it is not self-luminous. In missile guidance systems a part of the field, the target, is self-luminous, while the background light for the most part originates outside the field of view. The importance of the distinction between self-luminosity and non-self-luminosity is only secondary. More basic is the question of whether it is possible to control the nature of the radiation, either in the design of the equipment or during its use.

22.6.3.3 One important function of the optical system is the rejection of radiation from outside the field of view. The field stop in many instruments is an embodiment of this function. In other cases, as in certain types of scanning systems, the limitation of the field is obtained by more elaborate means.

22.6.3.4 The type of information to be obtained from the field depends on the application. In spectroscopy, the object is to obtain a measure of intensity as a function of wavelength, without regard to the portion of the field of view from which the radiation comes. In spectroscopic systems for on-stream process control, the object is, in addition, to present this information, or a portion of it, as a function of time. In image-forming systems, the object is to have rather detailed information concerning radiation intensity as a function of position throughout the field. If the image-forming system is of the "color-translation" type, at least some information about the spectral distribution of intensity at each point of the field must also be provided. In infrared search systems, the object is to know, from moment to moment, the presence and location within the field of small areas, or targets, having radiation characteristics slightly different from those of the remaining background portions of the field.

22.6.4 Detector characteristics.

22.6.4.1 The manner in which the information is obtained from the incoming radiation, and in fact, subject to the over all requirements for the instrument, the precise nature of the information, depends greatly on the kind of energy detector which can be used.

22.6.4.2 To the optical designer, the energy detector is the last surface of his system, which must receive and absorb all the useful energy collected by the system. The nature of the detector and its associated equipment imposes limitations on his choice in bringing this about, and also on the minimum of light-gathering power which he must build into the optical system. Several characteristics of the detector are worthy of discussion.

22.6.4.3 For the purposes of this discussion, the detector may be considered as a figurative "black box" with an input, the radiation, and an output, usually an electrical signal in infrared devices. Ordinarily it is more useful to consider the input as flux, or flux per unit area of the detector, rather than as total energy. The flux may be expressed in watts, or in some other unit of power. As will be seen in the following paragraph, it is frequently necessary to consider the distribution of the flux as a function of wavelength. The output is measured in appropriate units, in megohms for example if it is the change in electrical resistance of the cell due to the incident radiant power. The responsivity of the cell is the output per unit input; in our example the number of megohms change in resistance per watt of input. (The term sensitivity is sometimes used for what is here called responsivity, but the word sensitivity has also been employed for a number of other concepts, so its use will be avoided altogether in this discussion.)

22.6.4.4 Two qualifications must be put on this concept. In the first place, the output per unit input may depend on the wavelength of the radiation. Thus to characterize a detector adequately it is necessary to give its responsivity as a function of wavelength; and to predict its response, the distribution of the incident power as a function of wavelength must be known. As a class, the detectors known as photoelectric detectors are highly wavelength dependent. The thermal detectors as a class have substantially constant responsivity regardless of wavelength, and the spectral distribution of the radiation can be ignored.

22.6.4.5 The second qualification arises from the fact that the output may not be strictly a linear function of the input, even allowing for spectral effects. Many detectors show saturation effects when strongly irradiated. Frequently the detector is operated under conditions such that the response is substantially linear. In other cases the concept of responsivity must be modified suitably.

22.6.4.6 Another important characteristic of the cell is called its detectivity. It is a measure of the smallest input, or smallest change in input, that can be reliably detected. All detectors have a random output, not related to the input, known as noise. As a rule of thumb, the increment of input necessary to produce an increment of output equal to the noise may be taken as the minimum detectable input. When expressed as power, this is known as "noise-equivalent power". The larger the noise-equivalent power, the poorer is the detector for small inputs. The reciprocal of the noise-equivalent power is called the detectivity. (The word sensitivity has sometimes been used to mean the detectivity.) Detectivity depends on the type of detector, on the way it is made, and on the environmental and electrical conditions under which it is used. Ordinarily the detectivity is improved by keeping the detector area small.

22.6.4.7 When the input is suddenly changed, the output does not change instantaneously, but takes a finite time to adjust to the new level. The time constant is a measure of the time required for such an adjustment. It is important in predicting the response of the detector to short bursts of radiation, and in determining its suitability for use in scanning systems and in other systems in which the input is made to vary at high frequency. As a class, thermal detectors have much larger time constants than photodetectors.

22.6.4.8 In choosing the size of the detector it is to be remembered that flux density, rather than total flux on the detector, is the criterion of the amount of output which will be obtained. For example, if two square lead sulfide cells of similar characteristics be operated under similar conditions, but the area of one is twice the area of the other, then the outputs of the two will be equal when the flux per unit area on the two is the same, although the total flux on the larger is then twice that on the smaller. Sometimes a large cell may be operated under conditions not practical with a smaller one so as to produce a higher output at a given flux density (for example, a photocell of large area can safely be operated at a higher bias voltage than a small one), but in many cases this is not practical (for example, the voltage available for biasing may be limited) and flux density is the criterion of the output obtainable. In general this is an advantageous situation, since it is usually possible to use a more compact optical system to produce a given flux density on a small area than on a large one. Within limits, small detectors have better detectivity than large ones of the same kind. As a consequence in critical situations where detectivity is an important characteristic small cells are used. In photoconductors, dimensions of a few tenths of a millimeter are common.

22.6.4.9 The responsivity of a detector may depend somewhat on the way the flux is distributed over its surface. There may be local "hot spots" which are more responsive than the rest of the surface. Since the photoconductive cells are used with a bias voltage, the response depends on the way the radiation falls with respect to the points at which the leads make contact with the detector. For this reason it is desirable to plan the optical system so that the non-uniformity will not be a disadvantage. This ordinarily means that the detector is not placed at, or immediately adjacent to, an image plane, but rather at an image of the entrance aperture of the system.

22.6.5 Target detection and location.

22.6.5.1 An important class of problems is exemplified by airborne infrared search systems. In applications of this sort it is necessary to have a large field of view, and to detect and locate small, weak targets which are at great distances from the optical system.

22.6.5.2 There may be large amounts of radiation in the field besides that coming from the targets which it is desired to detect. Such radiation is called "background". The problem thus is one of distinguishing the radiation of the target from that of the background. It is necessary to take all possible advantage of differences between the target and the background. One important difference lies in spectral distribution, the target radiation usually having a peak wavelength different from that of the background. (The spectral distribution must be evaluated at the optical system. The intervening atmosphere is in effect a part of the field of view, and its absorbing and scattering characteristics must be taken into account.) Contrast between target and background is further increased by the use of optical filters to absorb as much as practical of the radiation at wavelengths at which the background is stronger than the target. (For a discussion of infrared filters see Ballard et al, loc. cit.) Choice of the type of detector depends in part on its having adequate responsivity in the spectral region near the peak wavelength of the target.

22.6.5.3 The technique known as spatial filtering is frequently used to take advantage of the dimensional differences between the target and other sources of radiation likely to be in the background. For a discussion of spatial filtering see, for example, Aroyan. *

22.6.5.4 Detection and location of the target within the field of view is accomplished by dividing the field into elements by some means and observing either the difference in flux between an element and adjacent ones, or the change in flux in each element with time. Since the intensity difference between target and background is small, the detector must be chosen and used so as to have good detectivity, and the optical system must have a large aperture to insure that the difference between target and background can be recognized by the detector.

22.6.5.5 An attractively simple scheme for providing the necessary subdivision of the field uses an objective which forms an image of the field at its focal plane. In the plane of the image is placed a rotating opaque plate carrying a set of small apertures so arranged that at any instant a single aperture is transmitting light from some small portion of the image, and during a single rotation of the plate the whole image is scanned. A condenser system placed behind the image plane collects the radiation and brings it to the detector. (The condenser is usually designed to form an image of the aperture of the objective on the detector.) The rotation of the plate can be related electrically to the output of the detector so that the system as a whole recognizes the portion of the field from which radiation is being transmitted at any instant. Attractive though it is, this simple scheme is rarely adequate because of the simultaneous requirements of large field and wide aperture. The inadequacy is not due entirely to lack of ingenuity on the part of the optical designer, but results from a fundamental limitation on the light-receiving ability of a small surface.

22.6.6 Receiving ability of a surface as a limiting factor.

22.6.6.1 The method used in the following analysis of the light-receiving ability of a surface is old, though it does not seem to be so well-known as desirable. See for example Drude. ** The method applies to any surface through which all the useful energy must pass, and thus its conclusions apply to focal surfaces as well as detector surfaces.

22.6.6.2 Suppose the instrument to be confronted by a black body, the surface of which is at least large enough to fill the whole field of view of the instrument. That is, it is large enough so that any ray which enters the optical system and passes into the surface whose light-receiving ability is being investigated can be considered to have originated in the surface of the black body. It is convenient though not necessary to assume that the black body is infinitely distant from the instrument. Assume that the black body is at some fixed, uniform temperature.

---

*Aroyan, G. F. "The Technique of Spatial Filtering." Proc. I.R.E. 47; 1561-68; Sept. 1959.

**Drude, Paul, "Lehrbuch der Optik," Leipzig, 1900. English trans., "The Theory of Optics," N.Y. and Dover, 1959

22.6.6.3 Suppose that the surface whose light-receiving ability is being investigated is the surface of a black body which is at the same temperature as that of the external black body. It follows from the second law of thermodynamics that, regardless of the nature of the optical system, the internal surface cannot receive from the external one an amount of flux greater than that which the internal surface itself is radiating, for otherwise the system would be acting as a self-operating heat pump. In most systems the flux received by the internal surface from the external one will be considerably less than the maximum, due to the finite aperture of the system, absorptions within the system, etc.

22.6.6.4 Let I be the number of watts per unit solid angle radiated by either surface per unit area of the surface in the direction normal to the surface. This quantity is determined by the black body temperature, and for our purposes is to be considered constant. Let A be the area of the surface whose light-receiving ability is being investigated, and n the index of refraction of the medium on that side of this surface on which the light is incident. (The detector surface may be exposed to air, for example, or it may be in optical contact with glass, light reaching the surface through the glass.) Let F_s be the total flux radiated by the surface into the medium with which it is in contact. Then it can be shown that

$$F_s = \pi I n^2 A \qquad (1)$$

(Drude, loc. cit.) Then F_s also represents the upper limit of the flux we can hope the surface would be able to receive from the external black body.

22.6.6.5 As an example of the significance of this result, suppose that a system is contemplated which has an objective aperture 150mm. in diameter, at the front of the system. It is desired to have the objective focus the energy from a 20° field (i.e., from 0° to 10° off-axis) on an image plane, and then to condense the light on a photocell. It is desired to determine the minimum focal plane area and minimum cell area. It is first necessary to write down the expression for the flux received at the aperture from the 20° field. For purposes of later discussion this formulation will be made more elaborate than would otherwise be necessary. Let

$$\bar{A} = \text{the area of the aperture} = 75^2 \pi \text{ sq. mm.}$$

22.6.6.6 It can be shown that the element of flux dF received at the aperture from that elemental portion of the external blackbody which lies in a direction making an angle α with the axis of the system (i.e., with the normal to the aperture) and which subtends a solid angle dw as seen from the aperture is

$$dF = I \bar{A} \cos \alpha \; dw.$$

Here I has the same value as in (1) since both the internal and the external blackbody are at the same temperature. The quantity $\bar{A} \cos \alpha$ is the projection of the area of the aperture in the direction of the source. From the point in the external blackbody on the axis of the system the aperture appears circular. From points farther away from the axis the aperture appears foreshortened, so that the effective area is only $A \cos \alpha$. Let

$$B(\alpha) = \bar{A} \cos \alpha$$

and name

$B(\alpha)$ the effective aperture function.

Then

$$dF = I \; B(\alpha) \; dw.$$

22.6.6.7 It is convenient to consider the elementary part of the external blackbody as being the annulus included between the two cones corresponding to field angles of α and $\alpha + d\alpha$ respectively. Then

$$dw = 2\pi \sin \alpha \; d\alpha$$

$$dF = 2\pi I B(\alpha) \sin \alpha \; d\alpha \qquad (2)$$

$$F = 2\pi I \int_0^{\alpha'} B(\alpha) \sin \alpha \; d\alpha . \qquad (3)$$

Here α' is the maximum value of the field angle from which radiation is to be collected.

22.6.6.8 In this example, since $B(\alpha)$ is known, (3) can of course be written

$$F = 2\pi I \bar{A} \int_0^{\alpha'} \cos\alpha \sin\alpha \, d\alpha .$$

$$= \pi I \bar{A} \sin^2 \alpha' , \tag{4}$$

or substituting for $\bar{A}$ and α',

$$F = \pi^2 I \, 75^2 \sin^2 10° . \tag{5}$$

The second law and equations (1) and (5) then require

$$F \leq F_s ,$$

whence

$$\pi^2 I \, 75^2 \sin^2 10° \leq \pi I n^2 A$$

and

$$A \geq 75^2 \pi \sin^2 10°/n^2 = 533/n^2 . \tag{6}$$

Thus if the focal surface must be in air, so that $n = 1$, we must expect its area to be greater than 533mm ^{2}. Similarly, if the detector be optically immersed in a medium having an index of refraction $n = 2.2$, its area must be greater than 110mm ^{2}. A detector having this area would have too poor a detectivity in many applications.

22.6.6.9 The treatment in the example can be generalized. In the example it was tacitly assumed that the system had rotational symmetry about the optic axis. Even if this is not the case it is still convenient to assume a set of spherical coordinates centered at some point in the optical system, using α for the polar angle and β for the azimuth angle. Consider an elementary portion of the external blackbody which lies in the direction (α, β) and subtends a solid angle dw as seen at the optical system. Consider a bundle of rays starting from the element of blackbody. When the rays reach the system, some will pass in and reach the focal plane and ultimately the detector. The others will be excluded by the various apertures of the system. The cross-sectional area of that portion of the bundle which does eventually reach the detector may be called the effective aperture of the system for the direction (α, β). We denote it by $B(\alpha, \beta)$, which may be called the effective aperture function of the system.

22.6.6.10 The flux collected by the system from the element of the blackbody is then

$$dF = I B(\alpha, \beta) \, dw \tag{7}$$

(we assume for the present that absorption and similar losses are negligible) and the total flux collected by the system is

$$F = I \int_\alpha \int_\beta B(\alpha, \beta) \, dw \tag{8}$$

the integral being taken over the whole field, i.e., over all directions (α, β) for which $B(\alpha, \beta) \neq 0$. It follows from the second law of thermodynamics, and from (1) and (8) that

$$\int_\alpha \int_\beta B(\alpha, \beta) \, dw \leq \pi n^2 A . \tag{9}$$

22.6.6.11 It is important to note that in (9) the radiation intensity, I, of the blackbody has cancelled out, and (9) is a condition on the characteristics of the system itself, regardless of the nature of the actual radiation field with which the system may be confronted. The expression on the right hand side may be taken as representing the maximum light-receiving ability of a surface of area A. The equation gives a fundamental limit to the combination of field coverage and aperture which can be achieved with a surface of area A.

22.6.6.12 When the system does have rotational symmetry we can proceed as in the example, considering $B(\alpha, \beta)$ to be a function of α only, and obtaining (3) which with (1) yields

$$2 \int_0^{\alpha'} B(\alpha) \sin\alpha \, d\alpha \leq n^2 A \tag{10}$$

22.6.7 Practical limits and techniques.

22.6.7.1 Of course practical difficulties prevent attaining the limiting light-receiving ability. To do so would require bringing in rays at all angles of incidence to the surface up to 90°. This is difficult to accomplish because of the precision required in constructing and focussing, and because of the resulting sizes and shapes required for the optical components. As a rule of thumb, something like 1/2 to 3/4 of the limiting light-receiving ability can be utilized. The former requires angles of incidence of 45° or greater at the surface; the latter, 60° or more. Consequently in practice one is limited to

$$\int_{\alpha} \int_{\beta} B(\alpha,\beta)\ dw \leqq k\pi n^2 A \qquad (9')$$

or

$$2 \int_{0}^{\alpha'} B(\alpha)\ \sin\alpha\ d\alpha \leqq k\, n^2 A \qquad (10')$$

wherein k may be 1/2 to 3/4. Losses by absorption and reflection in the system must of course also be taken into account.

22.6.7.2 As another example, suppose it is desired to monitor at 120° field, and that a detector area of 0.25mm^2 is chosen as having optimum detectivity under the expected operating conditions. Assume further that from knowledge of expected targets and backgrounds an aperture of at least 4400mm^2 is considered necessary to insure that the target will be recognized by the detector. (This aperture area is that of a circle 75mm in diameter). It is desired to use a simple scanning system of the type described in 22.6.5.5. The designer must decide whether he can do this.

22.6.7.3 It is evident at once that although in use the scanning plate permits only a small portion of the field to be viewed by the detector at any instant, nevertheless the system must be designed so that, if the plate were removed, the detector would view the whole field at once. That is, the light-receiving ability of the detector must be made to cover the whole field at the full aperture.

22.6.7.4 Without attempting to decide for the moment how it can be accomplished, assume that $B(\alpha)$ is to be made constant for the whole field of view, and equal to 4400mm^2. Then equation (10') becomes

$$8800 \int_{0}^{60^{\circ}} \sin\alpha\ d\alpha \leqq 0.25\, k\, n^2$$

or

$$8800\,(1 - \cos 60^{\circ}) \leqq 0.25\, k\, n^2 .$$

Assuming k may be taken as 3/4, this becomes $4400 \leq 0.1875\ n^2$.

22.6.7.5 If the cell is to operate in air, or a vacuum, $n = 1$, and the inequality obviously is not satisfied. The system as proposed cannot possibly meet the requirements. The scheme would have to be abandoned and another adopted which permits the detector to be employed in a system with adequate aperture but with a limited field.

22.6.7.6 There are various ways of doing this. A simple one consists essentially of building a system of adequate aperture and suitably small field, and then pointing it rapidly first at one portion of the field and then another. Thus each part of the large field is observed intermittently, although not continuously. The pointing is usually done by rotating mirrors. Another scheme causes the detector to scan the image at the focal plane of an objective by moving it about in the focal plane. The difference between this scheme and the one originally suggested lies in the fact that here the whole light-receiving ability of the cell is operative on the small portion of the field being viewed at the moment, while in the former only a small part was used at any instant, the rest being blocked off by the opaque portions of the rotating plate. Still other schemes use an array of cells to scan the image, or a mosaic of cells fixed in position in the image plane, or a combination of these methods. See Figure 22.3. A complete discussion of choice of an optimum system is too lengthy to be included here.

22.6.7.7 In equations (9), (10), (9') and (10') the square of the index of refraction, n^2, appears on the right hand side. This is indicative of the fact that the light-receiving ability of a detector surface is increased if it is in optical contact with some medium other than air. From analogy with oil-immersion microscope objectives, it is customary to say that the surface is immersed in the medium. The increase in light-receiving ability is analogous to the increase in numerical aperture of a microscope objective obtained by using an immersion system. The surfaces of the glass or crystal on which the cell is formed must be so disposed as to permit light from the field of view to strike the cell at all incidences from zero up to very high angles. If the substrate is simply a plane parallel plate in air, for example, the critical angle of refraction in the glass will limit the light-receiving power to that which the cell would have in air. If the cell is placed at the center of a hemisphere of the glass, however, the full light-receiving power can be used.

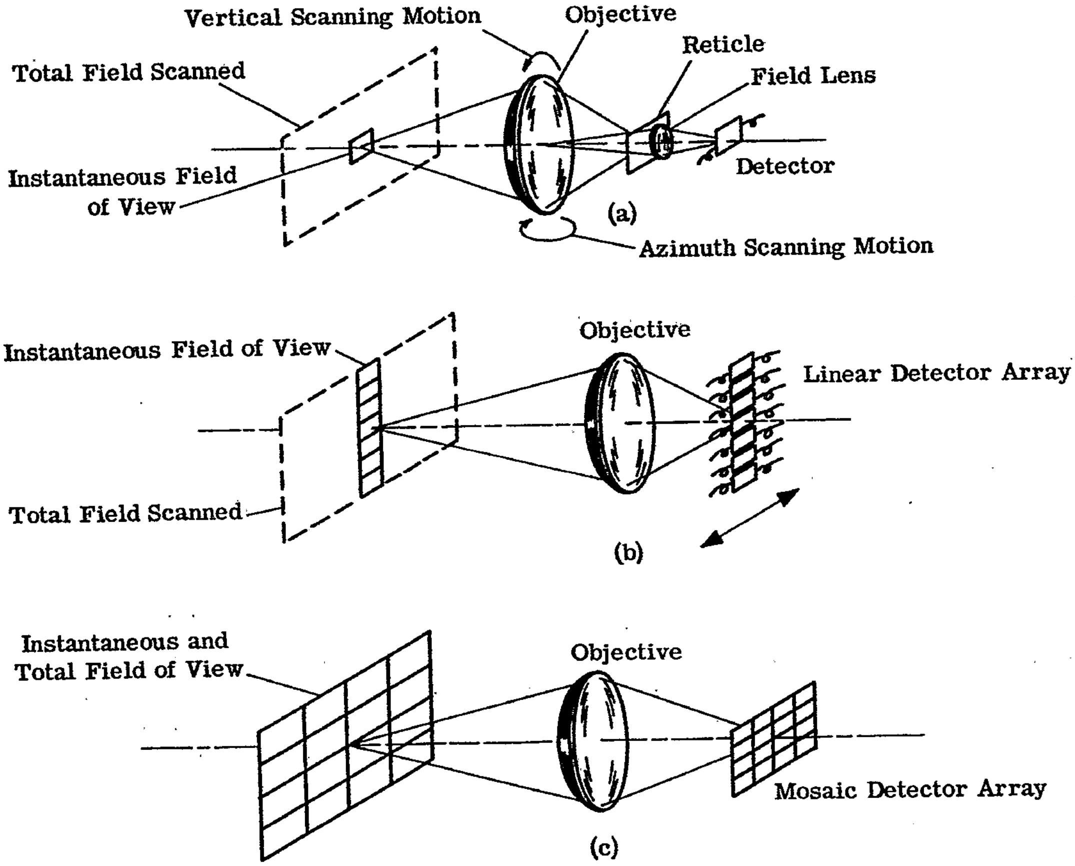
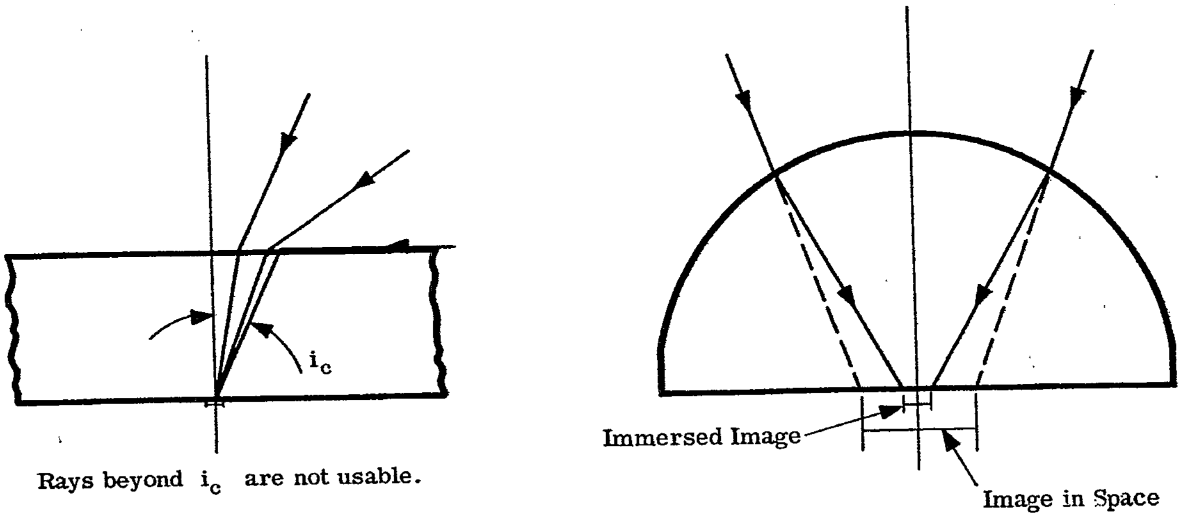


Figure 22.3 - Examples of basic scanning systems.



Rays beyond i_c are not usable.

Figure 22.4- Illustration of advantages of an immersed detector.

22.6.7.8 The benefits of cell immersion are often overlooked, and the design of a system is thereby made harder. However, as the possibilities become more widely appreciated, cell makers are giving more attention to the problems of producing immersed cells. These problems are difficult, because in addition to being a suitable material for supporting the cell, the substrate must also transmit the desired radiation. The higher the index, the greater the increase in light-gathering ability. Strontium titanate, for example, has an index of 2.2, and increases the light-receiving ability of the cell by a factor of 4.8.

22.7 SUMMARY AND CONCLUSION.

22.7.1 Advantages and disadvantages. In infrared work the designer may meet problems as varied and complex as those encountered in the visible part of the spectrum. Since the laws of reflection and refraction are the same at all wavelengths, the same basic design principles are used in both regions. The most important differences to which he must become accustomed arise from the natures of the available optical materials on the one hand, and from the requirements of some of the currently important infrared applications in the other. In addition he must remember that the resolving power obtainable with a given aperture is poorer than in the visible, due to the longer wavelength of the light. Out to about 2.7μ he may use ordinary glasses, but must allow for the fact that the dispersion characteristics of the several glass types are more nearly alike than in the visible. Beyond 2.7μ he must use materials whose characteristics may vary widely from those of optical glass, sometimes favorably and sometimes not. He will want to use reflecting systems more frequently than in the visible. The use of many infrared devices for military applications, particularly airborne ones, adds requirements of ruggedness, resistance to adverse environmental conditions, compactness and lightness of weight to the optical ones. Most such devices are part of instruments which are complex combinations of optics, mechanics and electronics, and the choice of the basic optical characteristics is only part of the process of choosing the optimum design parameters to meet the performance goals for the whole instrument. The designer needs to know enough about the characteristics of the whole instrument and the interrelationships of its parts to be able to contribute intelligently to the decisions in the choice of parameters. Especially, he needs to know something about the energy detectors used in the infrared, and how their limitations of responsitivity and detectivity limit his design.

23 MICROSCOPE OPTICS

23.1 INTRODUCTION

23.1.1 Scope. The material in this section will be devoted primarily to a discussion of the compound microscope, its characteristics, components, and various special purpose adaptations. However, in any discussion relating to visual instruments, the designer must keep in mind that the eye of the observer is an integral part of the optical combination, and that the degree of optical perfection in the human eye is as influential on the final retinal image, as is the degree of image perfection formed by the instrument's optical elements. The reader is urged therefore, to refer to Section 4 for a discussion of visual optics.

23.1.2 Functional relationships of microscope components.

23.1.2.1 The primary function of the high-power, compound microscope is to obtain information regarding the structure and optical characteristics of small specimens. This information is obtained by visually interpretating the manner in which the light transmitted by, or reflected from, the specimen is affected.

23.1.2.2 Usually, the specimen must be illuminated by intense artificial light, and it is only in rare and special cases that the specimen can be self-illuminated. However, the action of the specimen on the illumination system used may consist of absorption, reflection, diffraction, scattering, birefringence, or localized changes in the phase of the illuminating light waves. The purpose of the microscope then, is to form an image, based on the action of the specimen on the illuminating light waves, which can be interpreted in terms of the particular information with respect to the specimen, that is desired.

23.1.2.3 Since the human eye is only sensitive to color and intensity contrasts, the information derived from the image by the observer must be interpreted from these two effects.

23.1.2.4 The primary source of light can be of any number of high intensity light sources, however the light used must be concentrated on the specimen by a condenser system. The specimen affects the light as stated in paragraph 23.1.2.2, and the objective system of the microscope must be capable of receiving the altered light so that the maximum effects of diffraction, absorption, scattering, etc., may be transmitted by the objective and appear in the image as interpretable spacial variations.

23.1.2.5 In order to interpret the spacial variations in illumination, the objective must be capable of accepting and transmitting a wide angular beam of light, since the effects of the specimen on the light, as previously mentioned, especially diffraction, fan out from the specimen over wide angles. In the important case of diffraction, the more of the spectral orders the objective can receive, the more exact is the correspondence between the specimen and the structural details of the image.

23.1.2.6 Another requirement of the microscope objective is that the points and lines in a specimen be imaged sharply, so that the details in the image have a point by point correspondence with those in the specimen. This requirement necessitates a high degree of correction for aberrations.

23.1.2.7 All the available information regarding the specimen, as a result of its action on transmitted or reflected light, is contained in the primary image formed by the objective. As long as this information is contained in the primary image it is useless, since it must be interpreted in the brain of the observer. The simplest method for increasing the interpretability of the image is by means of magnification. By means of the eyepiece, the smallest significant details of the image can be resolved by the human eye.

23.1.2.8 There are other intermediate means available for interpreting the information from a specimen. The primary image may be magnified by projection, and formed on a photographic plate. The image could also be viewed by a television tube and an enlarged image presented on a screen.

23.1.2.9 To summarize then, the compound microscope is an instrument which transforms the action of a small object specimen on light waves into interpretable visual impressions, and in a broad sense any apparatus which accomplishes this function may be designated as a microscope.

23.2 CHARACTERISTICS

23.2.1 General. The compound microscope is characterized by the following requirements: high magnification (without a sacrifice of definition over a restricted size of field), a comparatively small true angular field, an illumination system, and resolution limited only by the wavelength of light and the numerical aperture of the objective. Similarly, it is desirable that the oblique aberrations be as well corrected as is consistent with the requirement for axial definition of the highest possible order. These characteristics determine the basic design of the compound microscope.

23.2.2 High magnification. Since the major function of the compound microscope is to view extremely small specimens, it must be capable of magnifying to such a degree that the smallest resolvable detail can also be resolved by the human eye. The highest useful magnification, expressed in diameters, is approximately a thousand times the numerical aperture of the objective used in the microscope. It should be noted however, that the compound microscope is not always used to view extremely small specimens and, in some instances, magnifications as low as 25 diameters are advantageous for viewing larger specimens.

23.2.3 True angular field. The true angular field of the microscope is, in most instances, small due to the following factors. The diameter of the primary image cannot be larger than that of the eyepiece, and present day eyepieces have become standardized in order to afford interchangeability with those of different manufacturers. The optical tube length, i.e., the image distance of the objective, in practice has become standardized, within limits, so that it is not difficult to interchange objectives made by different manufacturers, without significantly changing the magnification and correction of the objective. Since the true angular field, α, can be expressed as

$$\tan \frac{\alpha}{2} = \frac{y}{d} ,$$

where y is the half-diameter of the primary image, and d is the optical tube length, it may be seen that the true angular field has a maximum value in practice. The exceptions to this characteristic are the special microscopes which have extra large diameter eyepieces, and hence larger true angular fields. Actually, the size of the primary image is limited by the field diaphragm in the eyepiece, and this diaphragm must always be slightly smaller than the outside diameter of the eyepiece itself. In so-called negative type eyepieces, it is the virtual image of the diaphragm formed by the field lens which limits the primary image formed by the objective, but since the magnification of the field diaphragm by the field lens is not large, the statements above regarding field size are still applicable. The true angular field of the microscope may be considered to have a maximum value of less than 7°. For example, the half-diameter of the primary image field may be taken as not exceeding 10mm., and the optical tube length as 170mm. Substituting these values in the equation previously mentioned, it will be seen that the true angular field is 6°8'.

23.2.4 Illumination. The intensity of illumination in a compound microscope is a major factor due to the usually small size of the specimen being viewed, and because of the high magnification required to resolve the details of the specimen. As a result of the intense illumination required, light from an artificial source must be condensed onto the specimen. It is noted that in some cases, sunlight or skylight are used for illumination. Most specimens are thin and transmit light. For such specimens, the illumination falls on the back of the specimen, and the light is transmitted through the specimen into the microscope. For opaque specimens, the illumination is condensed upon the upper surface and only the light which is reflected from the specimen enters the microscope. This method of illumination is designated as vertical illumination. For most specimens however, transmitted illumination is used.

23.2.5 High resolution. High resolution is a basic requirement of the compound microscope, for it is upon this characteristic that the ability of the microscope to distinguish the fine details thereof, is based.

23.2.5.1 Factors determining resolving power. In compound microscopes, the source of light is most often an incandescent lamp provided with a condenser (in photomicrography and micro-projection, arc lamps are often used). The lamp condenser concentrates the light into a second condensing system, which is a part of the microscope proper and is known as the substage condenser. When vertical illumination is required, the substage condenser is not used, and other various forms of condensers are employed to condense light onto the specimen from above. Therefore, the resolving power of the microscope depends on the following factors:

(a) the size of the angle of the illuminating cone of rays passing through the specimen.

(b) the ability of the optical system to accept that which has been transmitted by the specimen and to transmit a wide cone of rays.

(c) the refractive index of the material between the specimen and the first surface of the optical system comprising the microscope's objective.

23.2.5.2 Limit of resolution. The concept of numerical aperture (N.A.) is essential in expressing the limit of resolution of the microscope. As illustrated in Figure 23.1, n is the refractive index of the medium in which the specimen, S, is immersed; θ is the half angle of the cone of incident rays; and the numerical aperture (N.A.) of the cone of rays is $n \sin \theta$. In a compound microscope, a glass cover slip, and in the case of immersion objectives a layer of immersion fluid, intervenes between the specimen and the entrant surface of the optical system. In this case, the numerical aperture becomes the product of the lesser index of refraction and the sine of the angle θ in that medium. The limit of resolution of the compound microscope, i.e., the least distance between two objects that can be seen as separate, is equal to the wavelength of light (λ) divided by the sum of the numerical apertures of the substage condenser and the microscope's objective lens.

23.3 COMPONENTS OF A COMPOUND MICROSCOPE

23.3.1 General. In order to realize the requirements for microscopic observations, the simplest form of an optical system is shown schematically in Figure 23.2. Figure 23.3 illustrates how these optical components (except the light source A and the lamp condenser B of Figure 23.2) are incorporated into a modern compound microscope. In Figure 23.3, the mirror below the substage condenser and the reflecting prism in the body between the objective and eyepiece are for mechanical convenience and are of no optical significance. The initial optical element of many compound microscopes is a mirror, which reflects light from the source into the remainder of the optical system. The mirror usually presents no design problems. One side is flat and reflects the light into the substage condenser. However, when extremely low powered objectives are used, the substage condenser usually will not illuminate the entire field of view because of its increased size. In this instance, the substage condenser may be removed, and the second side of the mirror, which is concave, will condense the light onto the specimen.

23.3.2 Illumination systems.

23.3.2.1 Simple illuminator. The simplest form of illumination for a compound microscope is a broad diffusing source, such as a ground glass, placed in front of an incandescent bulb. This source is imaged directly onto the specimen by means of the substage condenser (Figures 23.2 and 23.3). However, any granules in the ground glass would be visible in the field of view unless the image of the ground glass was slightly defocussed. Therefore, such a source is satisfactory only for low power microscopy, because of its lack of illumination for high

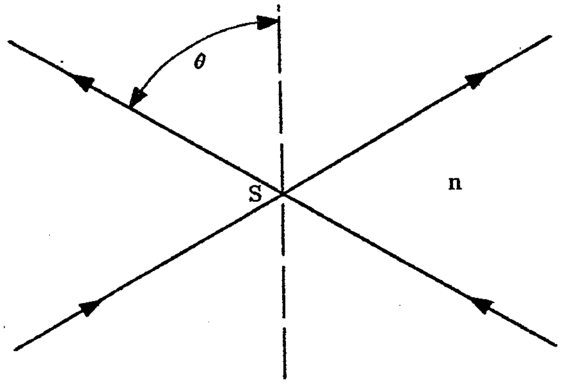


Figure 23.1- Determination of numerical aperture.

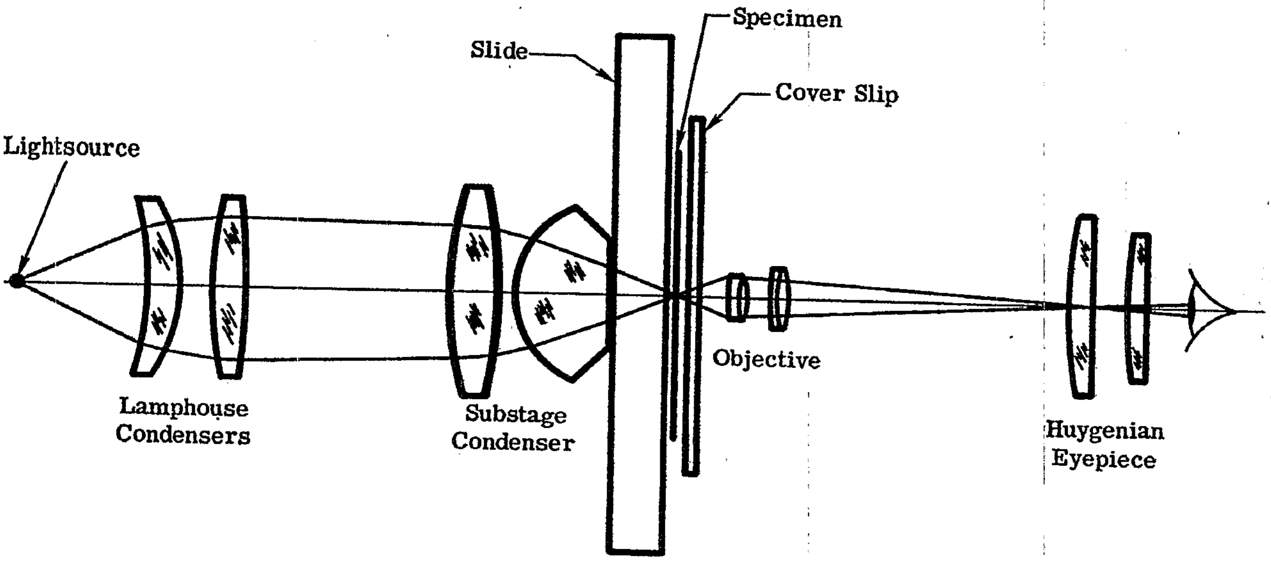


Figure 23. 2- Optical schematic of a compound microscope.

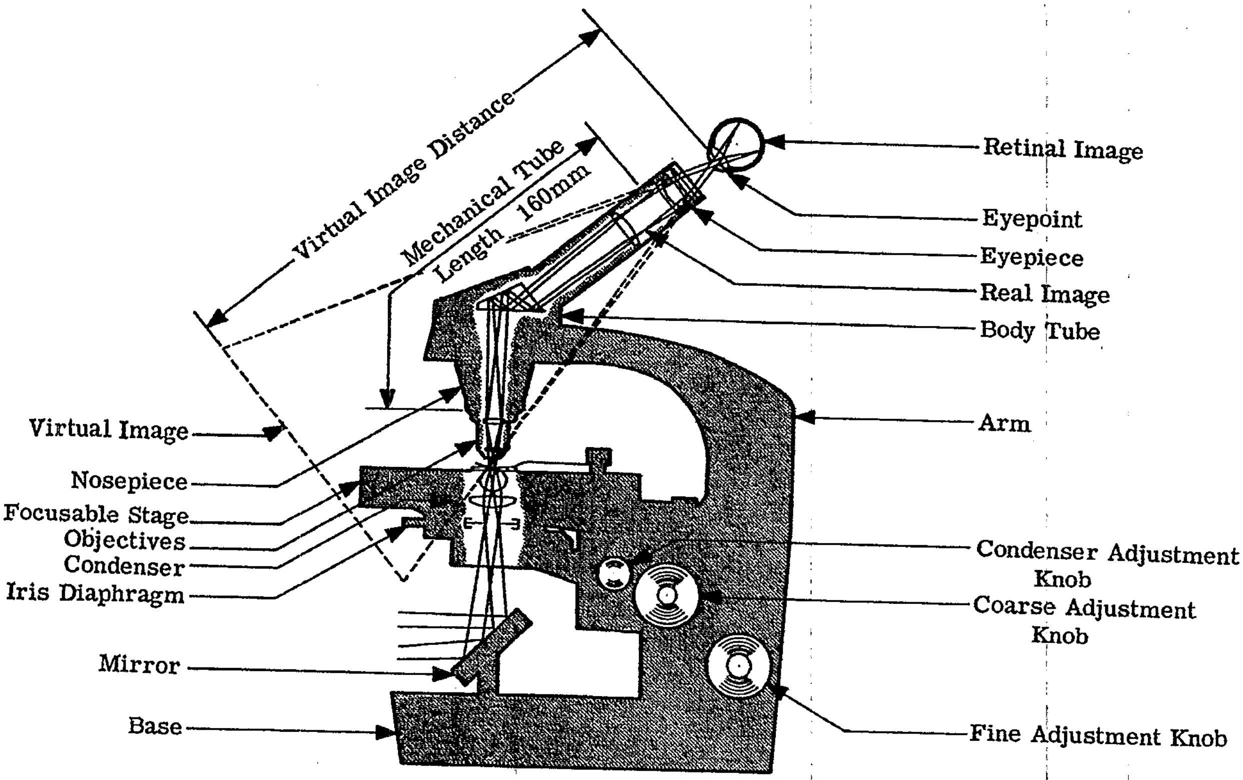


Figure 23. 3- Optical and mechanical features of the microscope.

power work. It should be noted that when the source of light is focussed directly onto the specimen, the illumination is designated as critical illumination. A more efficient microscope illuminator than that discussed here previously is shown in Figure 23.4 and consists of a monoplane filament lamp behind which is a spherical reflector and in front of which is a condenser--generally a two lens system. The filament of the lamp is near the center of curvature of the spherical reflector, and the reflected images of the strands of the filament are located between the strands themselves. This not only allows the reflected light to be utilized, but also forms a more nearly uniform primary source of light. The condenser in this system may be focussed so that the light source is imaged in the vicinity of a ground or opal glass, which in turn is focussed by the substage condenser onto the specimen (critical illumination).

23.3.2.2 From the preceeding paragraph it can be seen that critical illumination has the defect of not providing completely uniform illumination over the area of the specimen, and especially for photomicrography this is disadvantageous. A system known as Kohler illumination is used to overcome this difficulty. In this system, the lamp house condenser is used to focus the primary light source onto the substage iris diaphragm, placed at the front focal surface of the substage condenser, shown in Figure 23.4. Hence the light emerging from the substage condenser consists of parallel rays, which are re-imaged by the microscope's objective at its rear focal plane. The substage condenser now focuses the lamp house condenser, Figure 23.2, onto the specimen. Since the lamp house condenser is nearly uniformly illuminated by the light source, the field of the specimen is very uniformly illuminated. For cases in which the field must be uniformly illuminated, e.g., photomicrography, a form of Kohler illumination must be used unless the light source is very uniform as with a ribbon filament lamp.

23.3.2.3 Optical requirements for an illuminator.

23.3.2.3.1 The spherical mirror offers no design problem other than that of its aperture being large enough so that the reflected light will pass through the optical elements that follow.

23.3.2.3.2 The lamp house condenser should be large enough so that its image (formed by the substage condenser on the specimen--Kohler illumination) will fill the field of view. In addition, the focal length of the lamp house condenser should be correctly determined, in order that the particular primary light source will be large enough to fill the substage iris diaphragm.

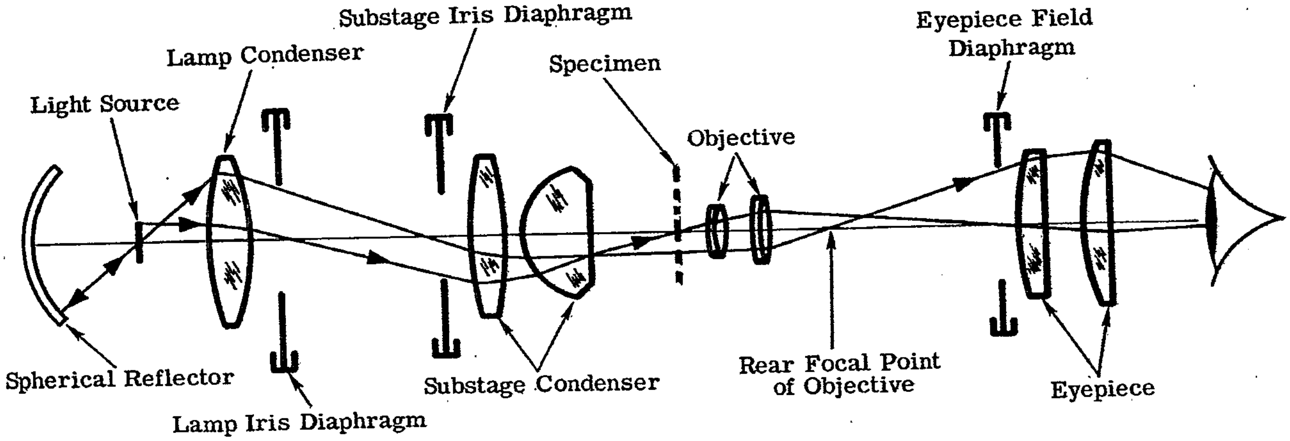


Figure 23.4- Kohler Illumination, schematic diagram.

23.3.2.3.3 An iris diaphragm is often located very close to the lamp condenser. With this design, the condenser (in Kohler illumination) and the iris diaphragm are imaged on the specimen, and if the iris diaphragm is adjustable in diameter, its image can be made to precisely fill the field. An adjustable iris diaphragm will prevent illumination of a greater area of the specimen than is necessary, and will also prevent scattered light, with its resultant loss of contrast, from entering the microscope.

23.3.2.3.4 The lamp condenser is usually a two lens, air-spaced system. It should be as well corrected for spherical aberration as is possible. The focal length must be correctly determined in order to image the filament of the lamp large enough to fill the iris diaphragm of the substage condenser (Kohler illumination) at a convenient distance (approximately 15 inches). The diameter of the condenser must be large enough so that its image, as formed by the substage condenser, covers the specimen field, when viewed with a 16mm focal length,(10x) microscope objective. It is readily apparent then, that the smaller the light source, the greater must be the speed of the condenser.

23.3.2.4 Vertical illuminators. For opaque specimens, vertical illumination is required for seeing surface details. Vertical illumination requires that the specimen field be uniformly and intensely illuminated, and that the illuminated field be limited to that portion of the specimen which is in the field of view. If the illuminated field is not limited as mentioned, an undesirable amount of light is scattered by the unviewed portion of the specimen, by the edges of the objective lenses, or by the walls of the objective. This scattering will reduce the contrast in the visual field.

23.3.2.4.1 Vertical illuminator, type A. In this type of vertical illuminator, as shown in Figure 23.5, the incident light is focussed on the specimen by being passed through the objective, in a reverse direction, and onto the specimen. The light source in this type is usually a low voltage, concentrated,filament bulb, located in a housing extending laterally to the axis of the microscope. The system consists of the light source, a condenser with an iris diaphragm mounted near the light source, a second condenser, and a semi-reflector at 45° to the microscope's axis for throwing light into the rear of the microscope's objective. The two condensers image the lamp filament at the exit pupil of the objective. The second condenser images the iris diaphragm (and the first condenser) at a virtual distance of about 160mm from the microscope's objective. Therefore, an image of the iris is formed by the microscope's objective on the specimen and it is uniformly illuminated. The field covered by the illuminated spot on the specimen can be regulated in size by adjusting the diameter of the field iris diaphragm, thereby preventing the scattering of light.

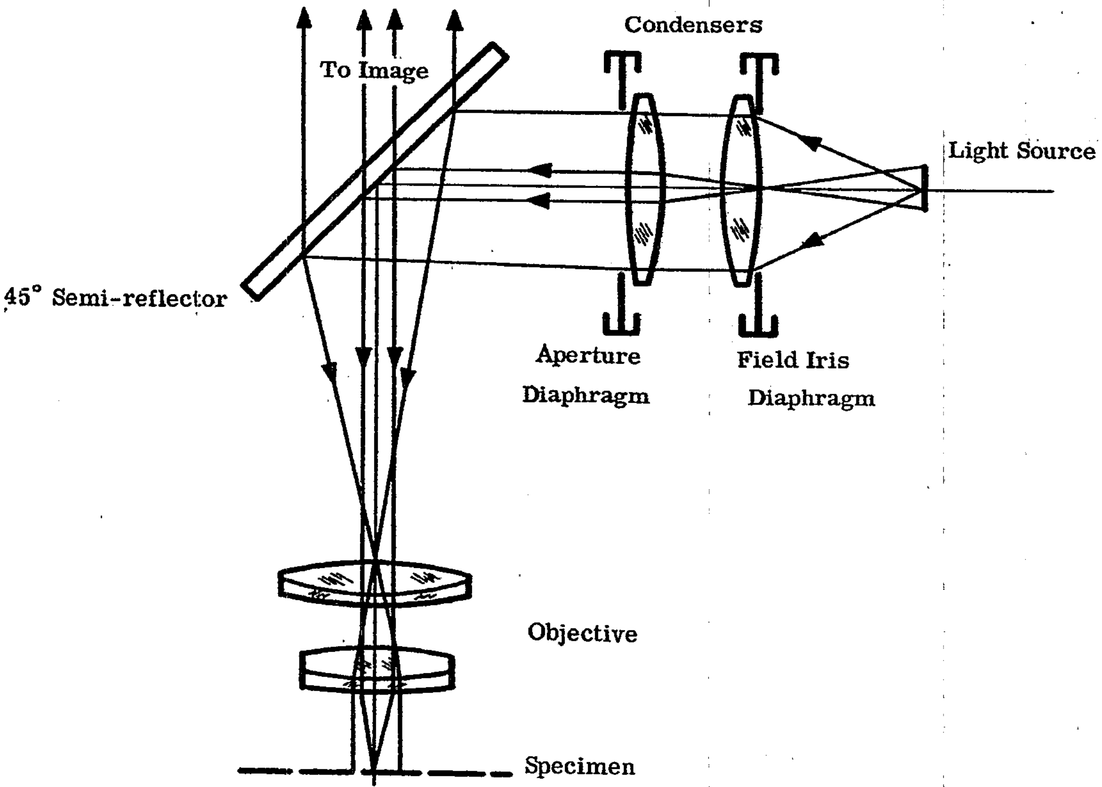


Figure 23.5- Vertical illumination.

23.3.2.4.2 Vertical illuminator, type B. In this type of illuminator, the light is not sent backwards through the objective, thereby preventing damaging internal reflections. This type of illuminator is best explained by picturing a condenser with a cylindrical hole through it axially, and the microscope's objective extending through this hole. In the actual design, the condensing elements surround the microscope's objective, and the illuminating cone of light is completely insulated from the objective's optics. Light is introduced into the condenser from the side through the use of an annular shaped full reflector. Such systems are called epi-condensers. Since the light must clear the lower rim of the microscope's objective and impinge on the specimen, it is obvious that the working distance from the objective to the specimen must be kept fairly large, therefore, this method of illumination is limited to the lower powers and numerical apertures of the objectives.

23.3.3 Substage condensers. The substage condenser concentrates light onto the specimen at a numerical aperture equal to that of the objective. There are two general classifications of condensers: the non-achromatic and the achromatic.

23.3.3.1 Non-achromatic. The non-achromatic substage condenser is used with achromatic microscope objectives which will be discussed in paragraph 23.3.4.3

23.3.3.1.1 The most common non-achromatic substage condenser is the Abbe condenser with a numerical aperture of 1.25. It is a two lens, air-spaced system, with a double convex lower lens and a plano-convex, hyperhemispherical upper element. When immersion fluids are used, the working distance between the upper plane surface and the image of the light source is equal to the optical thickness of the microscope's slide, plus a small space of a few tenths of a millimeter between the upper surface of the condenser and the slide. Reference to Figure 23.6 shows that no attempt is made to correct the chromatic aberration of the Abbe condenser, since both elements have positive power. The only variables available to the optical designer are the refractive indices of the elements, the shape of the first lens, and the separation of the two elements. The designer must adhere to a given working distance from the upper plano surface to the upper surface of the microscope slide, so that the light source (usually at a distance from infinity to a foot) can be focussed onto the specimen. In addition, the required numerical aperture is an additional consideration. A third requirement is that the focal length be such that the light source is imaged at a size sufficient to cover a specified area of the specimen, when the size and distance of the light source have been specified. (In practice, the light source is at least two inches in diameter at a minimum distance of one foot, and the field to be covered is that of a 16mm focus microscope objective). The requirements of numerical aperture and field coverage will determine the focal

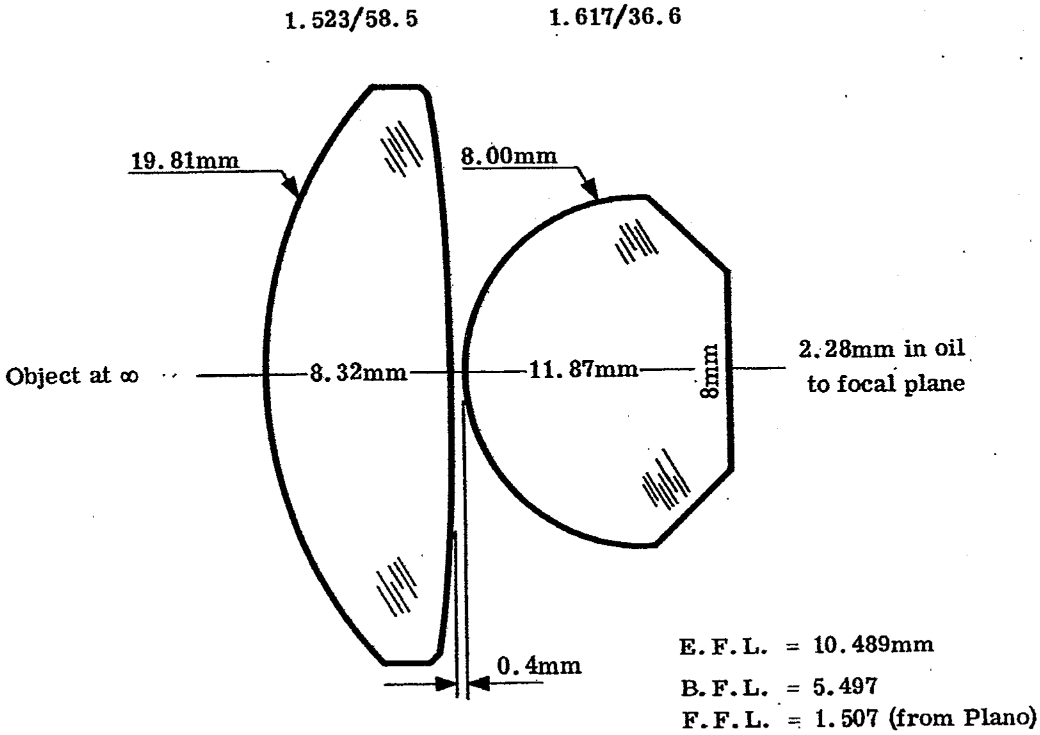


Figure 23.6- Optical layout of a 1.25 NA Abbe Condenser.

length of the front or first lens. The only aberration the designer may control in an Abbe condenser is spherical aberration. In the initial design of an Abbe condenser, graphical methods are useful and they are usually followed by mathematically triangulating a set of axial meridional rays through the system. In this way, the back lens may be bent to correct for spherical aberration. In conclusion, the two lens Abbe condenser is most commonly used with achromatic, rather than apochromatic objectives. For the latter, an achromatic condenser is used which more nearly approaches a microscope objective in form, construction, and correction. As the design principles of these condensers so closely approximates those of objectives, the reader is referred to paragraph 23.3.4.

23.3.3.2 Achromatic. Achromatic condensers are more complex than the Abbe condenser, and may consist of a triplet, doublet, meniscus, and front lens. This construction affords an opportunity for the correction of spherical aberration, coma, and chromatic aberration. This design will be seen to be essentially that of a microscope objective having the same numerical aperture, namely 1.30 and 1.40. Microscope objectives will be discussed in paragraph 23.3.4.3 through 23.3.4.5. Achromatic condensers find their most useful application when used in combination with apochromatic objectives.

23.3.4 Objectives.

23.3.4.1 Classification of objectives. Microscope objectives are classified as achromatic, semi-apochromatic, and apochromatic. If a microscope objective has been designed to correct for spherical aberration for one color of the spectrum, and for axial chromatic aberration for two colors, it is classified as an achromatic microscope objective. If the objective has been designed to correct for spherical aberration for two colors, and the axial chromatic aberration for three colors, it is classified as a apochromatic microscope objective. If the objective has been designed for correction between these two extremes, it is classified as a semi-apochromatic microscope objective.

23.3.4.2 Reasons for classification. With the magnification and resolving power (therefore the numerical aperture) corrected, the designer must take into account additional factors. Since the microscope is an instrument of almost fixed image distance, the magnification of the objective is almost proportional to its focal length. The image distance of the objective is not quite constant, since the corresponding fixed distance in the microscope is the mechanical distance from the mechanical shoulder of the objective, where it makes contact with the nosepiece, to the mechanical shoulder of the eyepiece, where it in turn makes contact with the upper end of the body tube. When the body tube of the microscope contains prisms or other optics for special purposes, the preceeding statement is no longer applicable. When the body tube contains prisms or other optics for special purposes, the optical tube length can be made the same as that of a microscope not having optical elements between the objective and the eyepiece, through the use of auxiliary compensating lenses. The distance between the front principal point of the objective and the specimen must be the same for a series of objectives, if these objectives are to be parfocal, i.e., no shift in focus should be required as a change of objectives is made. In order to meet this condition, the distance from the second principal point of the objective to the mechanical mounting shoulder, is certain to be different with the different powers of objectives, so that a really constant image distance cannot be obtained. In general, the numerical aperture of a microscope objective must be increased with magnification. However, since the difficulty of correcting aberrations increases rapidly with an increase in numerical aperture, the complexity of construction of the objective also increases. A cemented doublet is satisfactory for numerical apertures below 0.25, and focal lengths of 32 and 48mm. A 16mm focus having a numerical aperture of 0.25 requires two cemented doublets in the system.

23.3.4.3 Achromatic. The cost of microscope objectives depends on the complexity of their construction, and the cost of the optical materials used. These factors increase with the numerical aperture of the objective and the degree of correction required. For most routine examinations of biological or industrial materials, the moderate corrections and construction of the achromatic type objective are sufficient. This class of objective for higher powers are constructed of the following: a hemispherical or hyper-hemispherical lens known as the front; followed by a meniscus, the second front; a cemented component, the middle; and a cemented component, the back. Some of these components can be omitted from the lower numerical apertures. For example, a 16mm focus, 0.25 numerical aperture objective achromat has two cemented doublets; an 8mm focus, 0.50 numerical aperture objective achromat has a front, no second front, but a middle and back; a 4mm focus, 0.66 numerical aperture objective achromat has a front, second front, a middle, and a back. A 1.8mm focus, 1.25 numerical aperture achromat objective has a front, second front, cemented doublet middle, and a cemented doublet back.

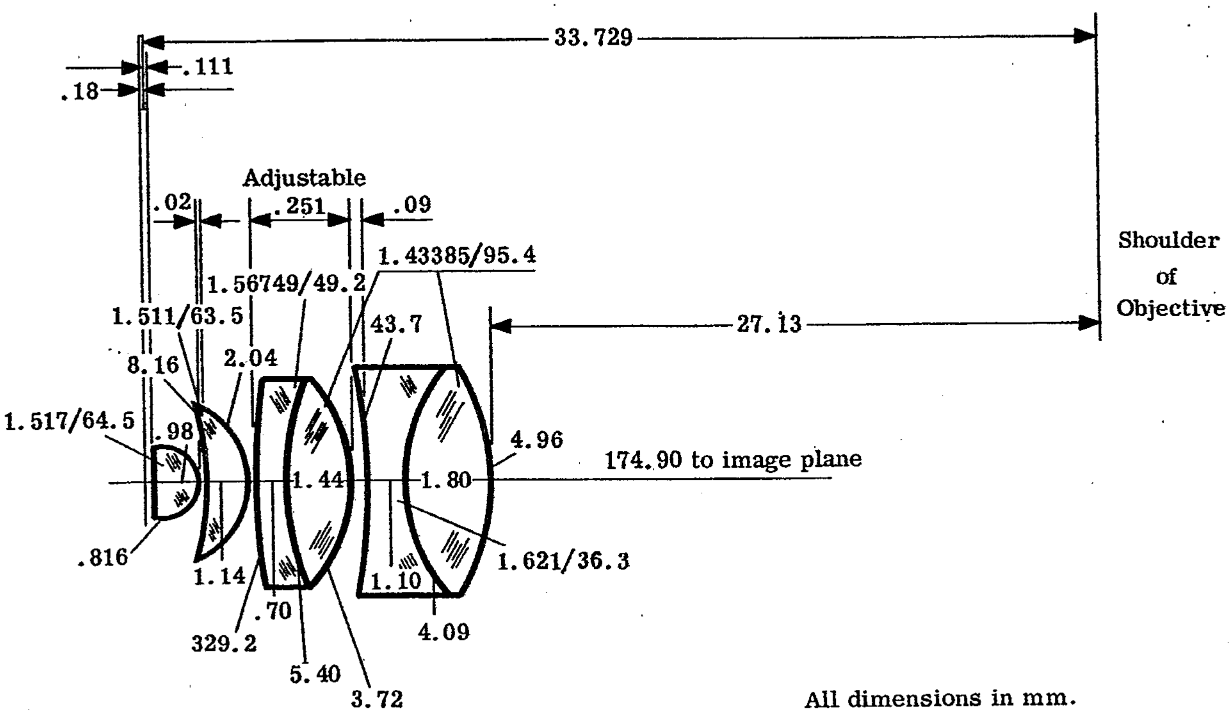
23.3.4.4 Semi-apochromatic. For more exacting routine microscopy, and for some kinds of research work, a higher degree of definition than that afforded by the achromatic objective is desirable. Semi-apochromats usually has flourite for one of its elements. Because flourite has a low refractive index, low dispersion, and a partial dispersion ratio different from glass, a better simultaneous correction for primary and secondary chromatic aberration and spherical aberration can be accomplished by its use as a positive element in a lens system. For example, if a flourite positive element is used with a flint glass negative element, a steep interface between the elements is attained, when the chromatic aberration is corrected. The over correction for spherical aberration, resulting from the steep interface and the large refractive difference at it, can be used to compensate for the under correction of other elements. By virtue of flourite's partial dispersion ratio being out of line with that of glass secondary chromatic aberration is favorably influenced. Constructional data for a semi-apochromat is shown in Figure 23.7

23.3.4.5 Apochromats. Apochromats are the most highly corrected of any of the microscope objectives. The optical design considerations involved are those described in 23.3.4.4, but they must be carried to the highest possible state of perfection. The correction is accomplished by the addition of optical components in the middle and back sections of the objective, and by the use of such crystals as alum and flourite to accomplish simultaneous correction for color, coma, and spherical aberration. Flint glass of the shortened spectrum type is also used in some constructions.

23.3.5 Eyepieces. There are three important considerations in the design of a microscope eyepiece. Since the object for the eyepiece is the image formed by the other elements of the optical system, the eyepiece can be designed to correct some of the residual defects in the other elements of the microscope's optical system. Also, the design of the eyepiece must be such, that a virtual image is formed anywhere between the point of most distinct vision (approximately 10 inches distant) and infinity. Finally, the eyepiece must be designed to correct lateral chromatic aberration.

23.3.5.1 Types of eyepieces. The main types of eyepieces used in compound microscopes are the Huygenian and Ramsden, and the type known as compensating eyepieces.

23.3.5.2 Huygenian eyepiece. For observational purposes, the Huygenian is often preferred to other eyepieces, since it can be completely freed of lateral color. The Huygenian eyepiece consists of two plano-convex lenses of the same type of glass (usually spectacle crown), with the field lens having a focal length approximately three times that of the eyelens, depending on the type of correction desired. The field lens and eyelens are separated in the body tube by a distance equal to twice the focal length of the eyelens. The combination is, therefore, free of lateral chromatism, and is most widely used with achromatic objectives (see paragraph 23.3.4). As is the case with all eyepieces, the limiting aperture for image forming bundles of rays is the exit pupil of the entire optical system of the microscope. The exit pupil is generally close to the second focal point of the eyepiece, and if a field stop or diaphragm is used, it should be positioned at the first focal point of the eyelens in order that its image will be formed at infinity. To some extent in microscopy, reticles are provided, and it follows that these should also lie in the plane of the first focal point of the eyelens. However, when such is the case the reticle is magnified by the eyelens alone, and even though the eyepiece combination as a whole



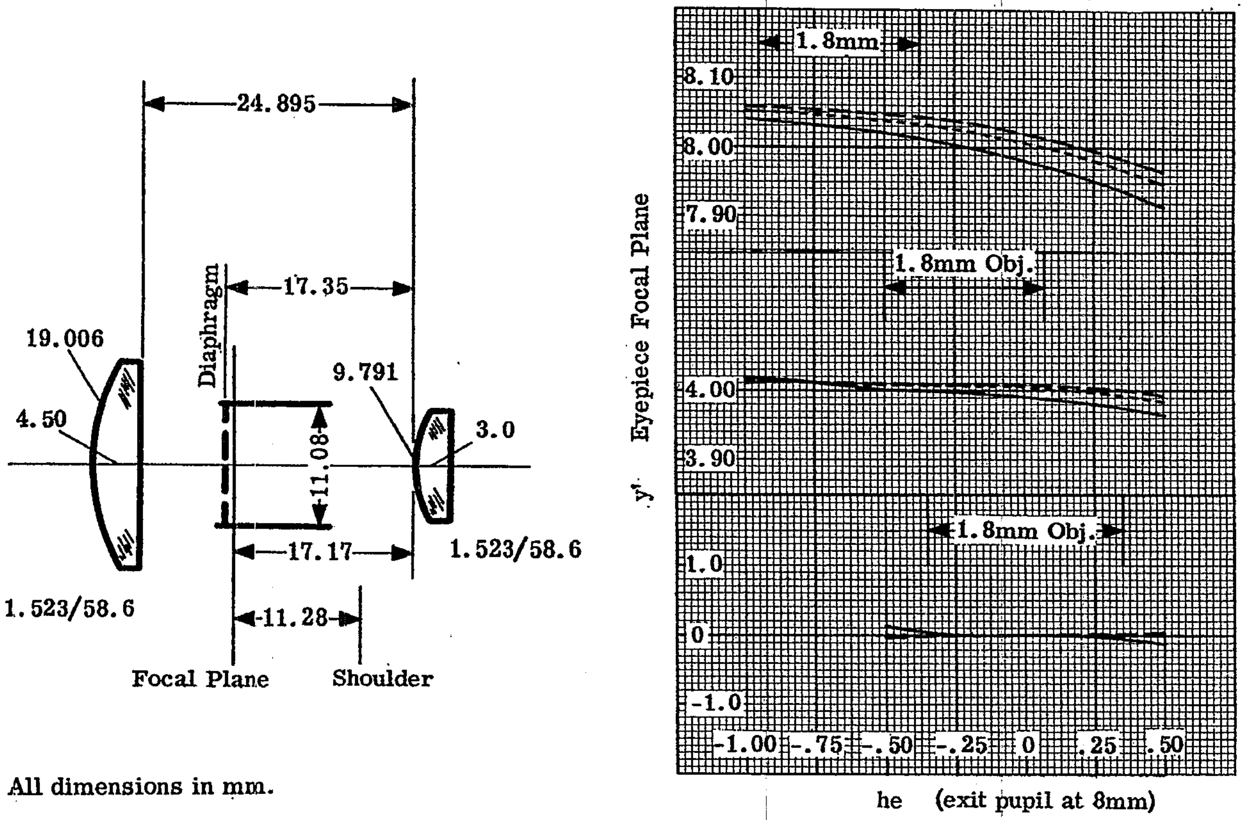
All dimensions in mm.

Figure 23.7- Optical layout of a 1.8mm flourite objective.

is free of lateral chromatic aberration, the corrections provided by the field lens are lacking and a large amount of aberration, particularly distortion and lateral color are introduced. To overcome this difficulty, the reticle used is kept so small that it is seen only at the center of the field. With respect to residual aberrations, the Huygenian eyepiece shows some spherical aberration, a large amount of longitudinal color, and marked pincushion distortion. An additional disadvantage occurs when this type of eyepiece has focal lengths less than one inch, since the eye relief is then usually too short for comfort. The reader is referenced to paragraph 6.11 where it is shown that a lens system such as this, can be designed to have constant equivalent focal lengths for all colors. An illustration and aberration graph of a Huygenian eyepiece is shown in Figure 23.8.

23.3.5.3 Ramsden eyepiece. A second type of eyepiece occasionally used with the microscope is the Ramsden eyepiece. The Ramsden eyepiece consists of two plano-convex lenses of the same type of glass (usually ordinary crown glass) and with equal focal lengths. The lenses are separated by a distance equal to two-thirds of the focal length of a single element. The focal point of this combination lies outside the system and so the eyepiece can be used to focus on an external reticle or cross hairs. With respect to aberration, the Ramsden eyepiece has more lateral color than the Huygenian, but the longitudinal color is only about half as great. The Ramsden eyepiece has about one-fifth the spherical aberration, and approximately half the distortion as found in the Huygenian eyepiece. The Ramsden eyepiece evidences no coma, and important advantage over the Huygenian is its 50 percent greater eye relief. An illustration of a Ramsden eyepiece is shown in Figure 23.9, and it is designated by usage as a positive eyepiece, in contradistinction to the negative Huygenian type.

23.3.5.4 Compensating eyepiece. The compensating eyepiece is used in conjunction with apochromatic objectives (paragraph 23.3.4.5) and as was previously stated, transverse chromatic aberration is a characteristic of these objectives. In order to correct for this aberration, an equal and opposite amount is introduced by the eyepiece. The eyepiece compensates the lateral color of the objective, and derives its name from this property. In addition to a definite amount of lateral color, the design of the eyepiece must correct for coma, spherical aberration and axial color, and its curvature of field and astigmatism must compensate those of the objective in so far as possible. In some cases, the observer wears spectacles, especially when the ocular defect is astigmatism (Myopia or hyperopia can be compensated by simply focussing the microscope), and it is therefore desirable to have the eyepoint of the microscope high enough so that there is sufficient space for the spectacle lenses between the back lens of the eyepiece and the eyepoint. This requires the back focal length of the eyepiece be sufficiently large in relation to the equivalent focal length, which determines the magnification. Such



All dimensions in mm.

Figure 23.8- Optical layout of a 10X Huygenian eyepiece.

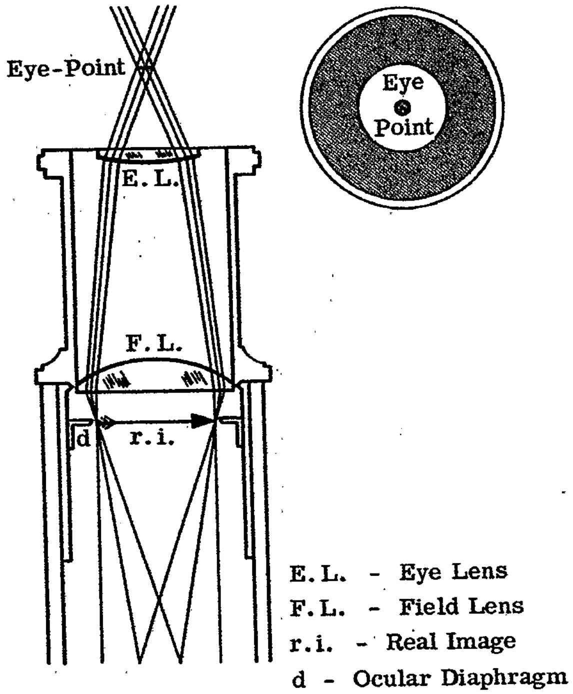


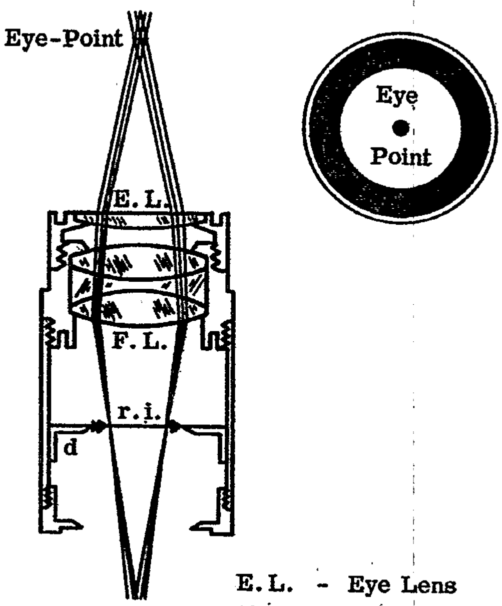
Figure 23.9- Optical schematic of a Ramsden eyepiece.

eyepieces are designated as "High-Eyepoint" and are illustrated in Figure 23.10. Compensating eyepieces can be of the positive or negative construction. Figure 23.11 shows the general construction of several powers of compensating eyepieces. The design of such eyepieces is dependent on the residual aberrations of the apochromatic objectives with which they are to be used. As shown in A and B of Figure 23.11, these eyepieces are of the negative type and are evolved from the Huygenian eyepiece by making the field lens and/or the eyelens cemented doublets, for purposes of correction. The high eyepoint compens is of the positive type and may be considered to be derived from the Ramsden eyepiece, by making the field lens a cemented triplet. The 30x compensator shown is essentially a ratioed form of the 10x High Eyepoint eyepiece. This construction prevents the eye distance from becoming too small, although the equivalent focal length necessarily is small in order to give the required relatively high eyepiece magnification.

23.4 DARKFIELD MICROSCOPY

23.4.1 General. In the ordinary microscope discussed previously, the illuminating bundles of rays enter the microscope objective and illuminate the entire field of view. The objects under examination are imaged as dark or colored details appearing against a bright background. Therefore, by this usual method of illumination, Brightfield Microscopy is accomplished. If the specimen is small, as for example with colloidal particles, or is practically transparent, the ordinary brightfield microscope does not offer sufficient contrast to render the objects visible. However, such particles have the property of scattering a portion of the incident radiation by means of diffraction, refraction, or reflection. In the field of darkfield microscopy, only the scattered light enters the microscope, while the direct illuminating beam entirely escapes the microscope's objective. Darkfield microscopy is accomplished by using the condenser to block the central portion of the light cone. The blocking of the entering light may be accomplished as detailed in 23.4.2 through 23.4.5. In both darkfield microscopy and ultra microscopy (paragraph 23.5) the objects appear to be self-luminous in a darkfield, and no light directly reaches the observer from an outside source. Light is only transmitted to the observer from the object being viewed.

23.4.2 Refracting darkfield condenser. A simple refracting darkfield condenser is an ordinary substage condenser provided with an opaque center stop which allows only rays traversing the outer zones of the condenser to be transmitted as shown in Figure 23.12. The effective numerical aperture of the microscope's objective must not be greater than the numerical aperture of the obscured central portion of the condenser, in order that the oblique hollow cone of rays transmitted by the condenser will not directly enter the objective. The oblique hollow cone of light will illuminate any object at its apex or focus, the object itself then deflects a part of this



E.L. - Eye Lens
F.L. - Field Lens
r.i. - Real Image
d - Ocular Diaphragm

Figure 23.10- Optical schematic of a high eye-point compensating eyepiece.

light into the microscope, and the object will then seem to be self-luminous in a dark field. The smaller numerical aperture of the illuminating bundle is about 0.7, while the upper limiting numerical aperture of the condenser is about 1.2. While such an illuminator is suitable for non-critical work, the refracting condenser has too much spherical and chromatic aberration for exacting darkfield use. In order to obtain a sufficiently dark background, it is important to have a very thin section of the specimen receive the focussed light. This condition will preclude any significant amount of aberration being present in the condenser. Darkfield condensers of the reflecting type may be well corrected for those defects and are generally used for high power work.

23.4.3 Reflecting darkfield condensers. The advantage offered by a reflecting darkfield condenser, with respect to the refracting type, is its ability to form a good ring of light for darkfield work, and its ability to minimize spherical and chromatic aberration in the transmitted bundle. More light will be scattered by the specimen if the difference between the inner and outer numerical apertures of this hollow cone is large. On the other hand, the microscope's objective is functioning at a numerical aperture not greater than the lesser numerical aperture of the hollow cone. This factor determines the amount of scattered light which can be used for image formation, and also determines the resolving power of the microscope. It is common practice to have the numerical apertures of the hollow cone cover the image from 0.7 to approximately 1.25. When objectives having numerical apertures greater than 0.7 are used, it is necessary to equip them with a funnel stop. The funnel stop will reduce the numerical aperture so that no direct light passes through the objective. For high power darkfield microscopy, not all the light can pass from the condenser to the specimen unless the specimen and its slide are in oil contact with the condenser. Some reflecting darkfield condensers are made with spherical surfaces or aspheric surfaces. Aspheric reflecting darkfield condensers are more difficult to fabricate, but are theoretically better corrected than the spherical type.

23.4.4 Aspheric darkfield condensers.

23.4.4.1 Paraboloid. A paraboloidal darkfield condenser is shown in Figure 23.13 (a). This condenser is a plano-convex block of glass with the reflecting surfaces forming a true parabola, at whose focus the specimen is positioned. Since the microscope's slide is in oil contact with the upper surface of the condenser, no aberrations are introduced.

23.4.4.2 Cardioid. In the cardioid darkfield condenser, the light rays undergo two reflections; one from the inner surface which is spherical, and one from the outer surface, which is cardioidal as shown in Figure 23.13 (b). This condenser, as is the case with the paraboloidal type, is free from chromatic and spherical aberration and, since it obeys the sine condition, is termed aplanatic. It is possible to observe particles as small as

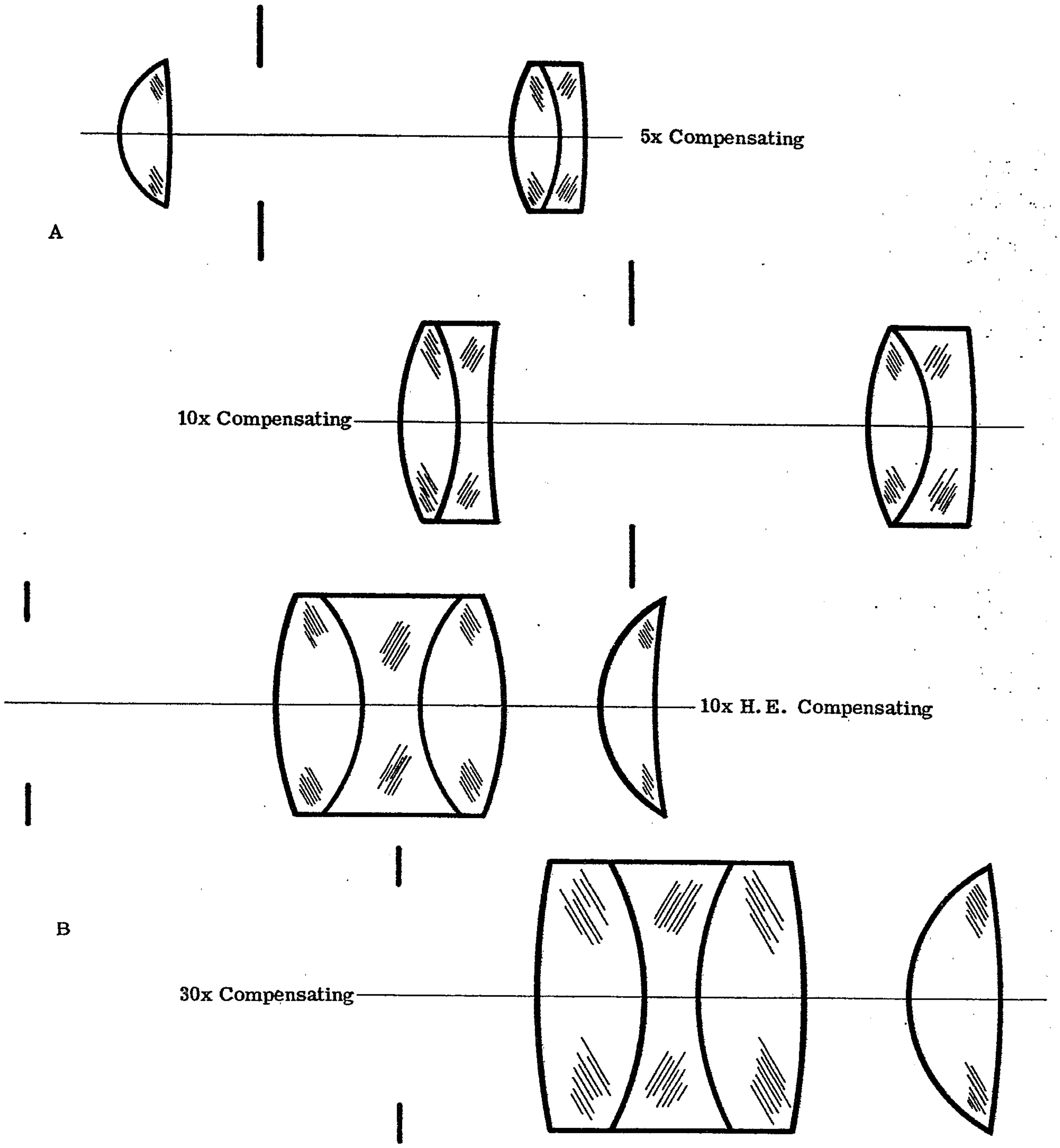


Figure 23.11- Typical compensating eyepieces.

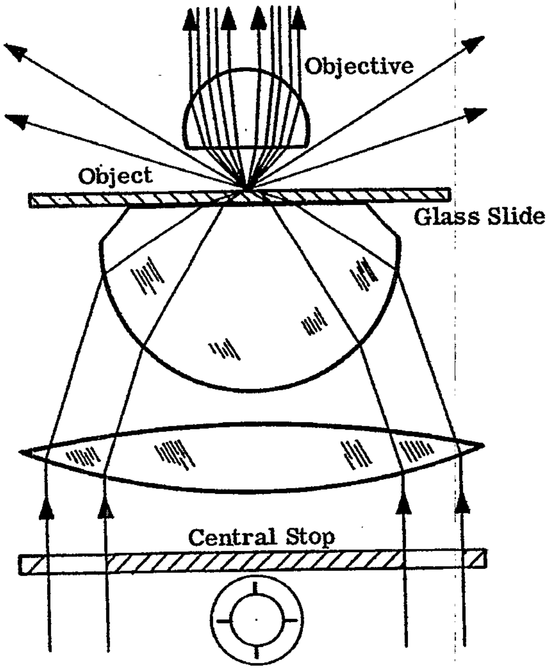
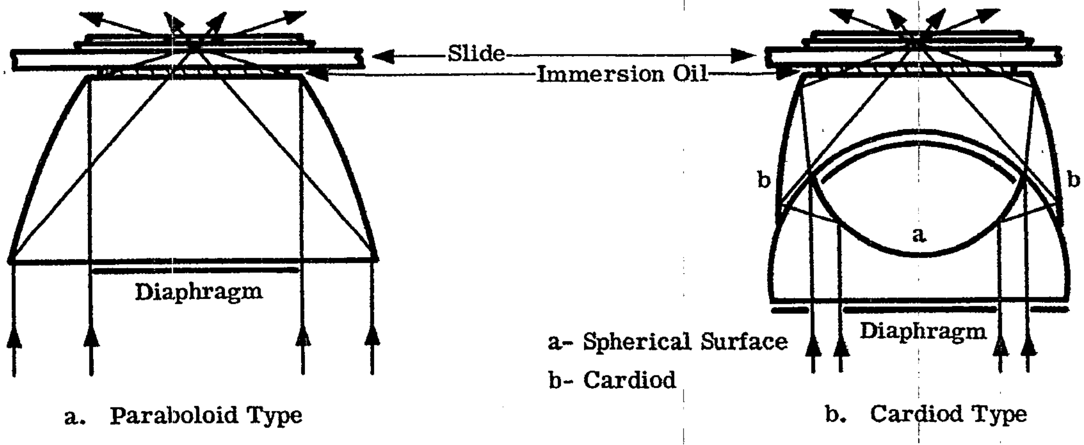


Figure 23.12- Refracting condenser with a central stop for dark-field illumination.



a. Paraboloid Type

b. Cardiod Type

Figure 23.13- Aspheric darkfield condensers.

0.000004mm in diameter under favorable conditions with this type condenser. The disadvantage of this type of condenser is the difficulty encountered in grinding and polishing a precise cardioidal surface.

23.4.5 Spherical darkfield condensers.

23.4.5.1 Bispheric. The bispheric darkfield condenser as shown in Figure 23.14 is constructed with both surfaces spherical, thereby avoiding the difficulty of precise grinding and polishing (as is the case with the cardioid type). The highly precise spherical surfaces can then be used with only slight deviations from theoretical considerations.

23.5 ULTRAMICROSCOPY

23.5.1 General. As indicated in the conclusion of paragraph 23.4.1, darkfield microscopy and ultramicroscopy are similar in their approach to the problem of studying objects or specimens. The two approaches differ only in the size of the object to be observed. Darkfield microscopes deal with objects of approximately 0.2μ or more in diameter, that is, those which come within the resolving power of the microscope. Ultramicroscopy deals with objects so small that the details cannot be resolved, but the presence of the object is inferred by the presence of light which the object transmits in the instrument. Some of the details of the specimen viewed with the darkfield microscope can be resolved, but some details are so small that they show simply as points of light, usually in the form of so-called diffraction discs. The larger details in the specimen come within the province of darkfield microscopy, while the smaller details are the concern of ultramicroscopy.

23.5.2 Characteristics.

23.5.2.1 Ultramicroscopes pass a narrow beam of light through the specimen at right angles to the axis of the viewing microscope. With a strong light source, such as a carbon arc, ultramicroscopes are excellent for viewing and counting particles in colloidal suspension. Figure 23.15 illustrates the essential components of a slit ultramicroscope. The arc (a) is imaged by the lens (b) on the cross slits (c). The cross slits (c) are imaged by a long working distance microscope objective (d) into cell (e), which is provided with two windows. The object to be viewed is introduced into the cell (e) and viewed by the microscope (f) through the upper window of the cell (e). In the cell, the Tyndall beam can be clearly seen in the microscope. By means of an eye-

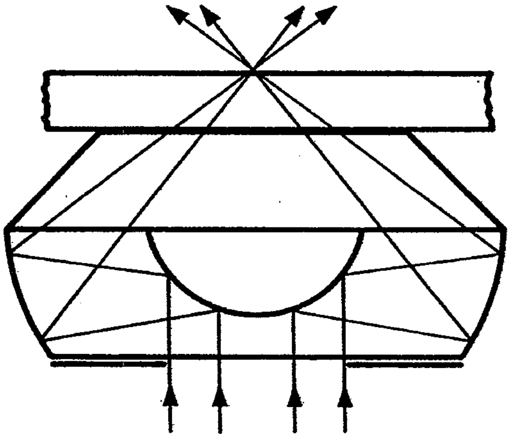


Figure 23.14- Bispheric (bicentric) darkfield condenser.

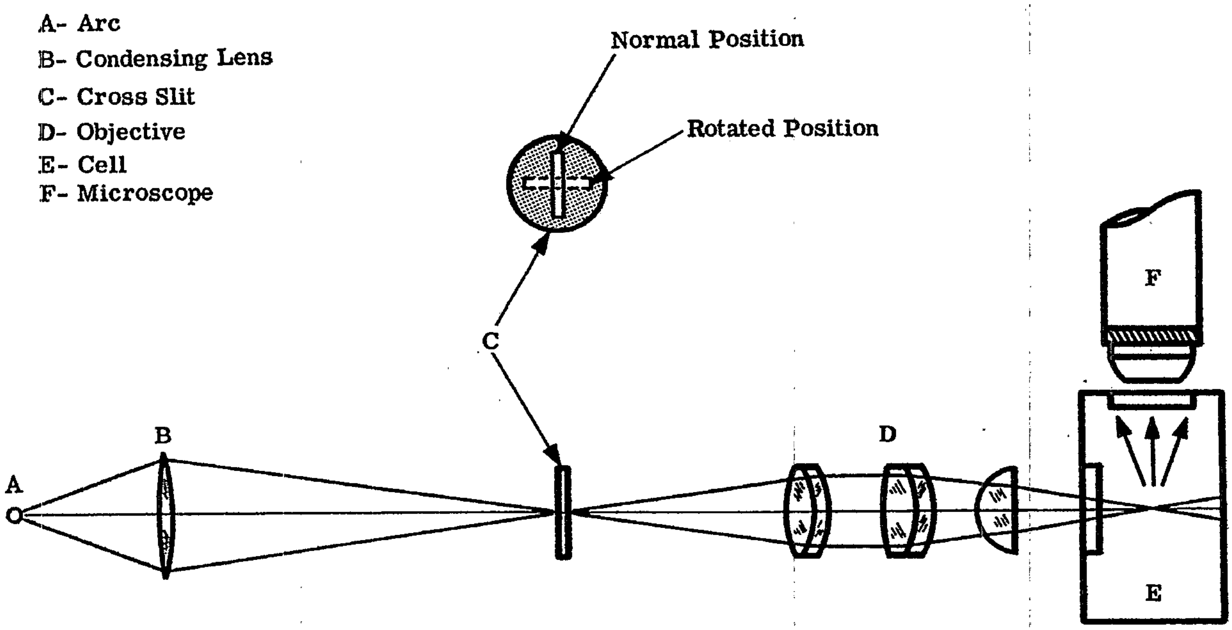


Figure 23.15-Elements of a slit ultramicroscope.

piece scale, the width and length of the beam can be measured. Therefore, if the cross slits (c) are rotated through 90°, the depth of the beam becomes the new width, and it can be measured. In this way, the volume of an illuminated portion of the contents of the cell (e) can be determined. In addition, the number of particles in the volume can be counted, and the number of colloidal particles per unit volume determined. Since the colloidal particles appear as diffraction discs, there is no need for high power or resolution in the viewing microscope.

23.6 PHASE MICROSCOPY

23.6.1 General. Nearly transparent materials having optical path (the product of the thickness and refractive index of the specimen) differences can be observed either with a phase or interference microscope. However, in contradistinction to the interference microscope (paragraph 23.7), the phase contrast is accomplished by the recombination in the image of direct light with the light deviated by the object after modification by a diffraction plate. It is interesting to note that a brightfield microscope may be converted to a phase microscope by the substitution of a phase condenser and phase objective. Similarly, the designer should keep in mind that a single contrast may be adequate for a given class of specimens, while other specimens may require several contrasts to reveal all of their structure.

23.6.2 Characteristics. The characteristics of a phase microscope can be seen from Figure 23.16. An annular diaphragm is placed in front of the condenser. When the annular diaphragm is uniformly illuminated, an image of it is formed in the objective near its focal plane, between the lens system. It can then be seen that all the light passes through this ring image or conjugate area when no specimen is present. However, when a specimen is being examined, some light is deviated through the rest of the area of the diffraction plate in the objective. The placing of a diffraction plate at this point differentially affects light deviated by the specimen, and the direct light from the background.

23.6.3 Principles.

23.6.3.1 The principles on which the phase microscope is based are shown in Figures 23.17 and 23.18. Figure 23.17 shows a light wave A' passing through a transparent object C. You will notice that A' has slowed down with respect to light wave A, which did not pass through the transparent object, and accordingly the two light waves are out-of-phase. However, the human eye and light-sensitive photographic plates are insensitive to phase differences, and as a result the image can scarcely be seen or photographed. Light wave A" in Figure

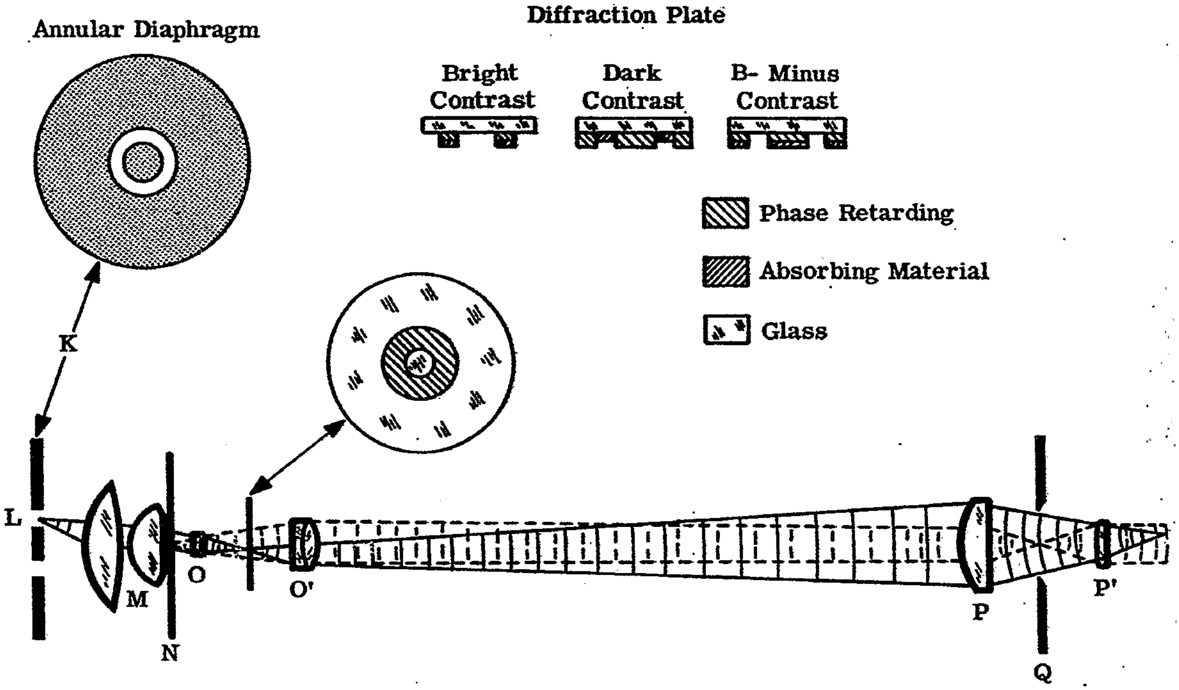


Figure 23.16- Elements of a phase microscope.

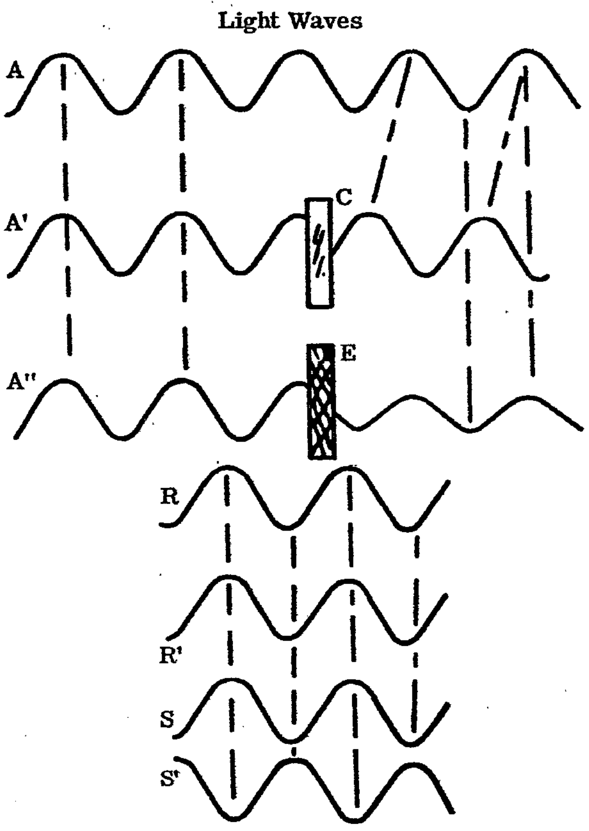


Figure 23.17- Passage of waves through mediums.

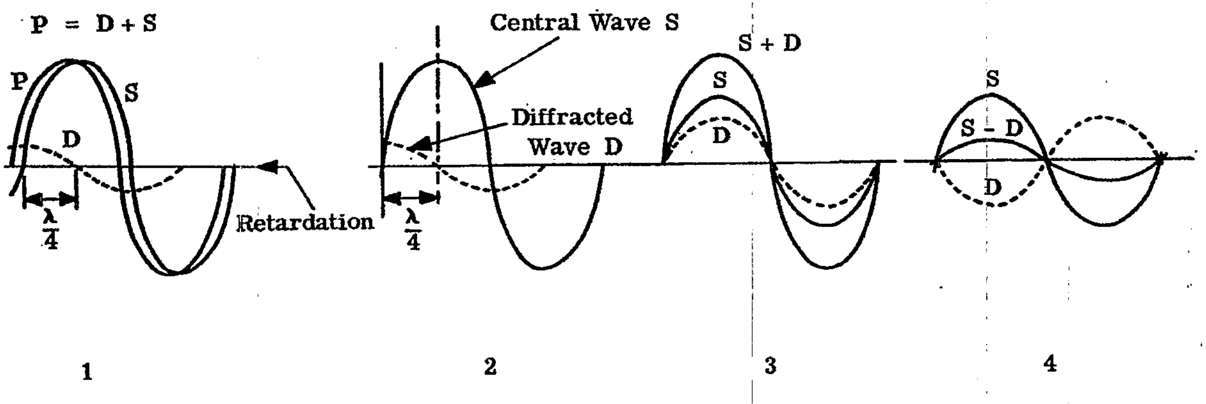


Figure 23.18 - Superimposed light waves.

23.17 passes through an absorbing medium E and is thereby reduced in amplitude as shown. However, in contradistinction to phase differences, amplitude differences are visible. When light waves of the same phase and amplitude are combined in the image as shown by R & R' in Figure 23.17, they add to produce brighter contrast. Similarly, dark contrast can be obtained by producing light waves which are out-of-phase or amplitude with each other as shown by S and S' in Figure 23.17, and combinations of amplitude and phase differences can be obtained which will produce lighter or darker greys.

23.6.3.2 Light waves may be superimposed as shown in Figure 23.18 Section 1 shows that the wave P, resulting from a slightly retarding particle, may be broken up into two waves S and D. The central wave S and the diffracted wave D, are shown again in Section 2. Section 3 demonstrates the result of using a bright contrast diffraction plate. The wave S has been partially absorbed, and wave D has been retarded, so that S and D are out-of-phase and produce a darker image.

23.6.3.3 The phase relationship of the light passing through a system of different optical paths has been altered, and the detail from the optical path differences, or slight absorptions of the specimen, will become visible within the microscope by the phase system elements (paragraph 23.6.2). By using an appropriate diffraction plate as previously discussed, it is then possible to increase or decrease the contrast of the image directly, or after reversing to change the contrast tone from bright to dark.

23.6.4 The diffraction plate. The diffraction plate consists of optical glass on which is evaporated, in vacuum, a very thin layer of metal, or a layer of a dielectric, or both. The layer of metal absorbs light, while the dielectric retards the light. The layer of metal or dielectric must be of sufficient size to cover either the image of the annular diaphragm formed in the objective or complimentary area of the remainder of the objective. These layers act upon the direct light from the background, and the deviated light from the specimen, so that recombination in the image will produce visible phase or absorption differences in the specimen.

23.6.4.1 The bright contrast diffraction plate absorbs, retards, or retards and absorbs the undeviated light, but has no effect on the deviated light. When this bright contrast diffraction plate is used, regions in the specimen of greater optical path will appear brighter than those of a lesser optical path.

23.6.4.2 The dark diffraction plate absorbs the undeviated light, and retards the light deviated by the specimen. The regions of greater optical path difference in the specimen will then appear darker. The effect of the dark diffraction plate is solely on the deviated light, and the degree of contrast is controlled by the width of the annulus and the thickness of the absorbing and retarding layer of the diffraction plate.

23.6.5 Disadvantages encountered with phase microscopy. As noted previously, phase contrast is accomplished in the phase microscope by the recombination of the direct and deviated light in the image, after diffraction. However, optical path differences and small absorption differences may be involved in this recombination and the resultant might be more appropriately termed "densiphase contrast" as suggested by Bennett, et. al. (1) Presently, phase microscopes have been modified to provide variable contrast, but not to measure the densiphase detail. Also, as the phase microscope redistributes the light in the image, haloes are often seen around the observed details, although the proper diffraction plate may lessen this condition. In addition, phase microscopes will make the optical path differences visible, but not their numerical magnitude.

23.7 INTERFERENCE MICROSCOPY

23.7.1 General. Interferometry, while well established in other fields, has only recently been applied to microscopy. Two methods of interference microscopy presently exist, the multiple beam method which is used extensively in the examination of surfaces of opaque materials having good reflection and in the examination of transparent materials, and the two beam method.

23.7.2 Characteristics.

23.7.2.1 Interference contrast. Interference contrast is accomplished by the recombination in the image of two beams of coherent light (from the same source), one of which is modified by passing through the specimen. In contradistinction to the phase microscope, the interference microscope will not produce haloes around the details. In addition, the interference microscope provides variable color contrast with white light illumination, and intensity variation in the color of the monochromatic light when monochromatic light is used. Similarly, with monochromatic light the interference microscope can provide measurement of the optical path differences in the specimen. It is interesting to note that when the thickness of the specimen is known, the refractive index can be measured, and in the case where the specimen is placed in a media of a different known refractive index, both the thickness and index can be measured. Also, interference microscopes have increased vertical resolution, but have the same lateral resolution as other light microscopes.

23.7.2.2 Multiple beam interference microscope. In the multiple beam method, the specimen to be examined is mounted between two flat, metalized, reflecting surfaces, and illuminated with parallel, monochromatic light. The recombinations resulting from repeated reflections of the light through the specimen produce fringes, which are used to measure the optical path differences (within reasonably transparent specimens).

23 7.2.3 Two beam interference microscope. With the two beam method, coherent illumination (from a single source) is so divided that part of the light passes in focus through the specimen, and the remainder passes to one side or is out-of-focus at the specimen. On recombining the light, the beams interfere to produce measureable patterns from which the optical path differences can be determined. This beam separation can be accomplished by reflection or polarization.

23.7.3 Principles.

23.7.3.1 The A O Baker interference microscope, Figure 23.19, illustrates the principles of interference microscopy. This microscope is fundamentally a polarizing microscope modified into a two-beam interferometer. The condenser has a birefringent plate which divides the light into two beams and the objective has a corresponding plate which recombines the beams after one of them has passed through the specimen. Above the objective is a quarter-wave compensator and an analyzer. Various eyepieces may be used to obtain different magnifications with the Shearing or Double Focus types of 10X, 40X and 100X objectives.

23.7.3.1.1 The polarizer below the condenser polarizes the light in a plane at 45° to the axis of the birefringent plate. The birefringent plate at the top of the condenser separates the polarized light into two beams which are plane-polarized at right angles to each other. One beam passes through the specimen, and the other passes to one side of the specimen in the Shearing system. In the Double Focus system one beam focuses at the specimen and the other spreads around the specimen to focus above it. The phase of the beam passing through the specimen is changed by the local variations in optical thickness in each portion of the specimen; while the changes in the reference beam depend on the average optical thickness of the specimen and the region around it in the Double Focus system; or the region to one side of the specimen in the Shearing system, as shown in Figure 23.20

---

(1) Bennett, A. H. Jupnik, H., Osterburg, H. and O. W. Richards, Phase Microscopy, pg. 11, John Wiley & Sons, New York, 1951.

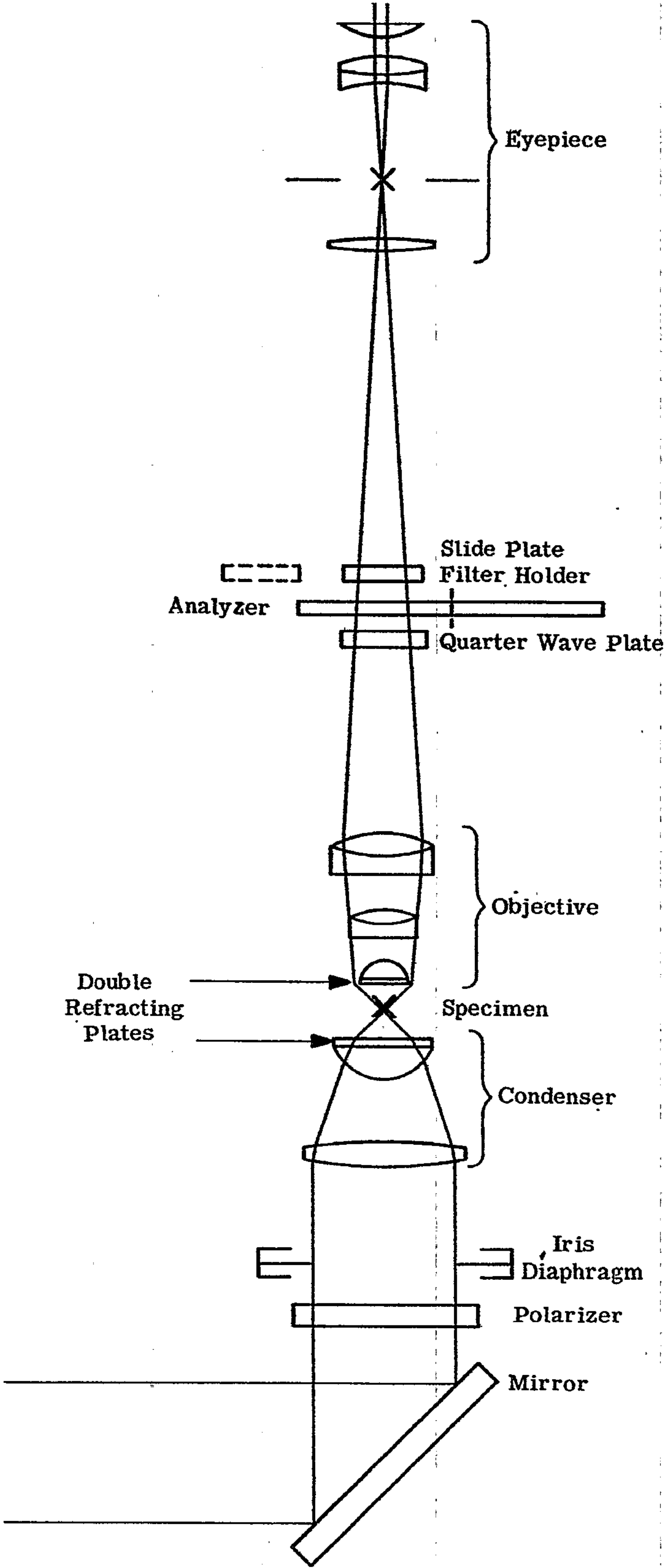
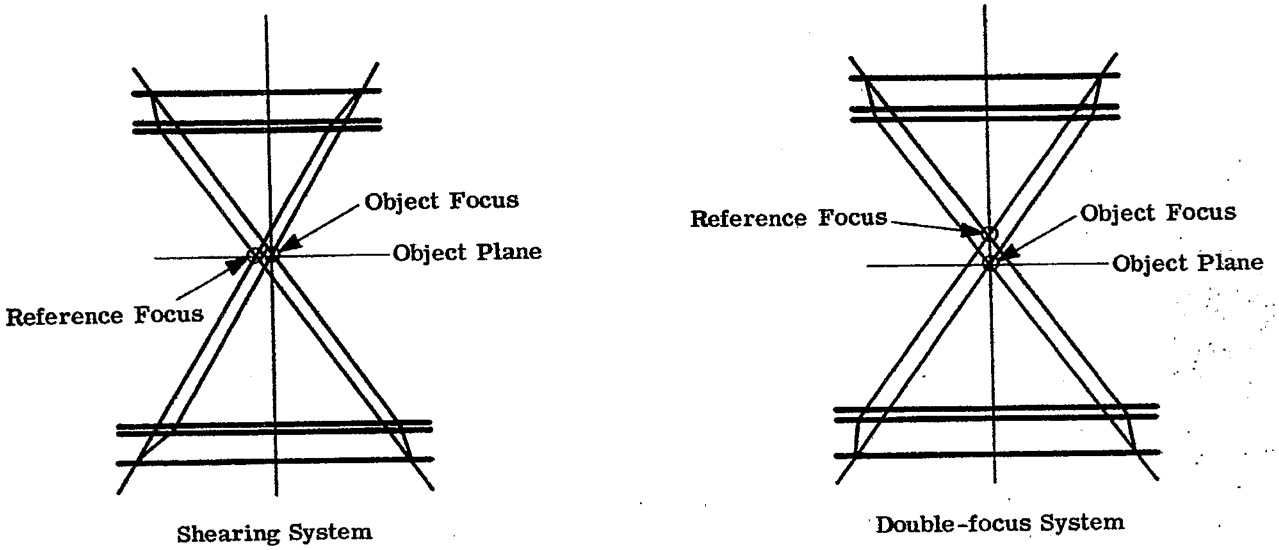


Figure 23.19- Optical schematic of AO Baker interference microscope.



Shearing System

Double-focus System

Figure 23. 20- Path of light rays through a Shearing and Double-Focus system.

23.7.3.1.2 The birefringent plate on the front of the objective unites the two beams and the quarter-wave plate changes the two oppositely polarized beams into left and right-hand circularly polarized light. The resultant of two circularly polarized beams is plane polarized light, the direction of the plane depending on the phase difference between the circularly polarized beams. Thus the phase differences in the specimen can be determined by turning the analyzer to the position of minimum luminance, or extinction, in the image.

23.7.3.1.3 A vector theory (2, 3) and an integration theory (4) have been proposed for the mathematical analysis of this type of microscope.

23.7.4 Interference colors.

23.7.4.1 Light passes through an optically denser medium more slowly than in a less dense medium and is retarded, with respect to light, through the less dense medium. The amount of retardation (phase difference) is proportional to the difference in refractive index for the particular wavelength considered.

23.7.4.2 For example, should the denser regions of a specimen illuminated with light from a mercury arc retard the blue light exactly one-half wavelength and the analyzer be set to extinguish the blue light, the specimen then would be seen only in the remaining yellow-green light.

23.7.4.3 With tungsten light, the blue is not limited to such a single wavelength as the mercury arc, but is a band of light (±440 - 490μ). A single particle cannot retard exactly to a half wavelength all of these blue wavelengths, therefore some blue will be lost and some transmitted and the particle will appear, more or less yellow, depending on the amount of blue lost.

23.7.4.4 Phase changes affect other colors in a similar manner and the actual interference colors depend on the composition of the light from the illuminator and on how the optical paths in the specimen retard or advance each wavelength (color). The relative amount of each wavelength passing through the analyzer determines the color of the particle.

---

(2) J. Roy, A Vector Theory of Phase Contrast and Interference Contrast. Micr. Soc. 75:23-37, 1955.

(3) G. Oster, A. W. Pollister, Physical Techniques in Biological Research 29-90 Vol. III, Academic Press, New York.

(4) ibid (3), 310-437, Vol. I.

23.7.4.5 Light is radiation to which the eye is sensitive (380-740Mμ) and can be seen. Interference microscopy is possible with invisible, infrared and ultraviolet radiation with receptors sensitive to the radiation used.

23.7.5 Illumination and filters.

23.7.5.1 Measurement of the optical path requires the use of monochromatic light to avoid the color interferences mentioned in paragraph 23.7.4. Monochromatic light is light of a single wavelength and is usually obtained from a single line of a spectral source. As the mercury arc has most of its radiation concentrated in a few lines it is the usual source for such illumination. The mercury arc has the further advantage in that the green line of 548mμ wavelength is quite close to the maximum sensitivity of the human eye (555mμ). Sodium light is suitable, although not as comfortable for visual use.

23.7.5.2 To isolate the light from a single line in the mercury spectrum a filter is used that transmits the light from the desired line and absorbs the light from the other bright lines in the spectrum of this source.

23.7.5.3 Some filters for use with mercury arcs are listed in Table 23.1. The least expensive is the Wratten 62 or 74, but these transmit only about 10% of the green light and do not exclude the light from the yellow line. The Corning CS4-120 transmits more of the green and no yellow. The Wratten 77 and 77A also transmit more green light than the 62 or 74, but for monochromatic light need to be combined with the 58 filter which reduces the light correspondingly. Filters for isolating the blue and yellow mercury lines are included in the table.

23.7.5.4 For some measurement, where the highest precision is not required, approximately monochromatic light is adequate and may be obtained with tungsten lamps and "narrow band" filters or with "interference" filters. The latter often have the disadvantage of low transmission.

23.7.5.5 The H85-C3 or H100-A4 (formerly AH3 and AH4) mercury arcs are satisfactory for many visual applications, but require long exposure when photomicrographs are to be made. More intense mercury arc sources such as the B-T-H 250 and the Osram HB0200 give more light, especially when monochromatic light is used, and are desirable for photography.

23.7.5.6 Light from the mercury arc without a color filter can be used for variable color contrast microscopy

Mercury line	Blue, 0.436μ		Green, 0.546μ		Yellow, 0.577μ	
Relative energy	80%		100%		88%	
Eye Relative luminosity (100 at 0.555μ)	1.8		98.4		89.8	
Filter	% Trans.	# Rel. Vis.*	% Trans.	# Rel. Vis.*	% Trans.	# Rel. Vis.*
Corning CS4-120	0	-	44	43	0	-
Corning CS-584	22	3	0	--	0	-
Ilford 625	0	-	35	34	8	6
Wratten 50	6.4	0.9	0	0	0	0
Wratten 62	0	-	10	9.8	0.5	0.4
Wratten 74	0	-	10	9.8	0.2	0.2
Wratten 77A	0	-	68	67	0	0
Wratten 77A 58	0	-	29	28	0	0
Wratten 77	0	-	74	73	0.5	0.4
Wratten 77 58	0	-	31	31	0.06	-
Wratten 58	0	-	42	31	11	9
Wratten 22	0	-	0	-	71	62

*Relative luminosity - Relative energy - filter transmission - relative luminosity of the ICI Standard Observer.

#Trans. = transmittance.

Table 23.1- Table of visual efficiency of isolating filters for the H100-A4 Mercury Arc (based on nominal values from manufacturer's literature).

although it will be seen that the mercury light has very little orange and red as compared to tungsten or daylight. When used with a filter the path differences in the specimen are seen only in the color of the filter, but with variable intensity.

23.8 POLARIZING MICROSCOPE

23.8.1 General.

23.8.1.1 The polarizing microscope is a brightfield microscope modified for examining the specimen in polarized light, with auxiliary equipment for measuring the effect of the specimen on the polarized light. A laboratory microscope can be used with polarized light by placing discs of Polaroid under the condenser and in or on the ocular. Such simple polarization will reveal the colors in birefringent materials and show strain.

23.8.1.2 For measurement, a specialized microscope is necessary. The polarizing microscope has a polarizing prism of the Nicol or Ahrens type under the condenser. The chemical type has a cap analyzer over the eyepiece and the petrographic type has the analyzer in a slide so that it can be pushed into or out of the optical axis in the body tube of the microscope. The upper lenses of the condenser are arranged so that they may be moved into or away from the optical axis of the microscope.

23.8.2 Characteristics.

23.8.2.1 Strain-free optics are necessary in the design of a polarizing microscope, and the objectives are usually mounted in centering rings of a quick change type of nosepiece.

23.8.2.2 When a Bertrand lens is pushed into the optical axis, it forms, with the ocular, a telescope for viewing the back aperture of the objective. A slot is provided for moving a quartz wedge or other compensators into the optical axis.

23.8.2.3 The ocular contains cross hairs and is positioned in the ocular tube to prevent rotation. The polarizer is rotatable to position it at 180° to the polarization angle of the analyzer and a centering rotatable stage with graduated scale and vernier are used to measure the orientation of the specimen.

23.8.2.4 The improved Polaroid is satisfactory and is used to replace the expensive crystal polarizers in some modern instruments and many special compensators, multiaxis stages and other auxiliary equipment are available and useful. The birefringence of biological materials is small, and more elaborate polarizing microscopes have been built to meet this need. One marked improvement is the rectifier for compensating depolarization from the curved objective lens that makes possible the use of the nearly full aperture of the oil immersion objective.

23.9 FLUORESCENCE MICROSCOPES

23.9.1 General.

23.9.1.1 Fluorescence microscopy can be accomplished with a brightfield microscope when the specimen is irradiated with ultraviolet radiation. A source of filtered radiation (usually a high pressure mercury arc) is necessary and an ultraviolet absorbing filter is placed on, or in, the ocular to prevent ultraviolet radiation not absorbed by the specimen from reaching the eye. A front-surface,aluminized mirror is more efficient than a silvered mirror.

23.9.2 Characteristics.

23.9.2.1 For short wavelength ultraviolet, necessary in the study of some minerals, the condenser and slide must be of quartz or other UV transmitting materials, or a catoptric condenser be used. Long wavelengths ($>330 m\mu$) UV pass through the ordinary microscope optics and they are satisfactory.

23.9.2.2 The most efficient system uses a crossed filter technic with a brightfield condenser. The lamp filter passes the radiation absorbed by the specimen and the ocular protective filter is chosen to absorb the ultraviolet, but to pass the light emitted by the specimen. When an efficient cross-filter system is not possible, a darkfield condenser is used with a thinner UV isolating filter. A colorless UV filter is usually required for the ocular as some of the UV may be scattered by the specimen into the objective.

23.9.2.3 Achromatic objectives and Abbe condensers are preferable as the chromatically corrected ones often contain fluorescent materials which introduce glare and reduces visibility. The Abbe NA 1.40 will concentrate more energy on the specimen than the usual NA 1.25 condenser.

23.10 THE STEREOSCOPIC MICROSCOPE

23.10.1 General.

23.10.1.1 The bi-objective, binocular microscope, reinvented by Greenough is made by combining two microscopes, Figure 23.21, so that the right eye sees with the right hand side, and the left eye with the left hand side. As each eye receives a separate disparate view, true stereopsis occurs. When the angles of the objective and ocular convergence are the same, true or orthostereopsis is provided. By changing these angles increased or decreased depth can be provided.

23.10.2 Characteristics.

23.10.2.1 Prisms are included to erect the image and such instruments are useful for dissection and for the examination of small parts.

23.10.2.2 Since two objectives are required, the mechanical limitations of placement limits the numerical aperture to about 0.12 and there is no advantage in using magnifications over about 120X.

23.10.2.3 A recent modification places the paired objectives in a rotatable turret with a single, large, corrected lens between them and the specimen. By turning the turret, magnification can be readily varied within the limitations of the cycloptic microscope. Another improvement is to build the paired objectives into a zoom system so that the magnification can be varied continuously throughout its range.

23.11 PETROGRAPHIC MICROSCOPE

23.11.1 General.

23.11.1.1 The optical system of the petrographic microscope has been so adapted that the methods of petrographic measurements can be made.

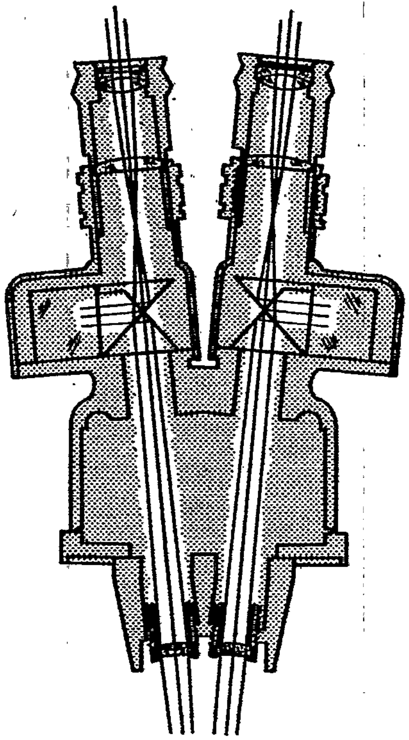


Figure 23.21- Optical schematic of a stereoscopic microscope.

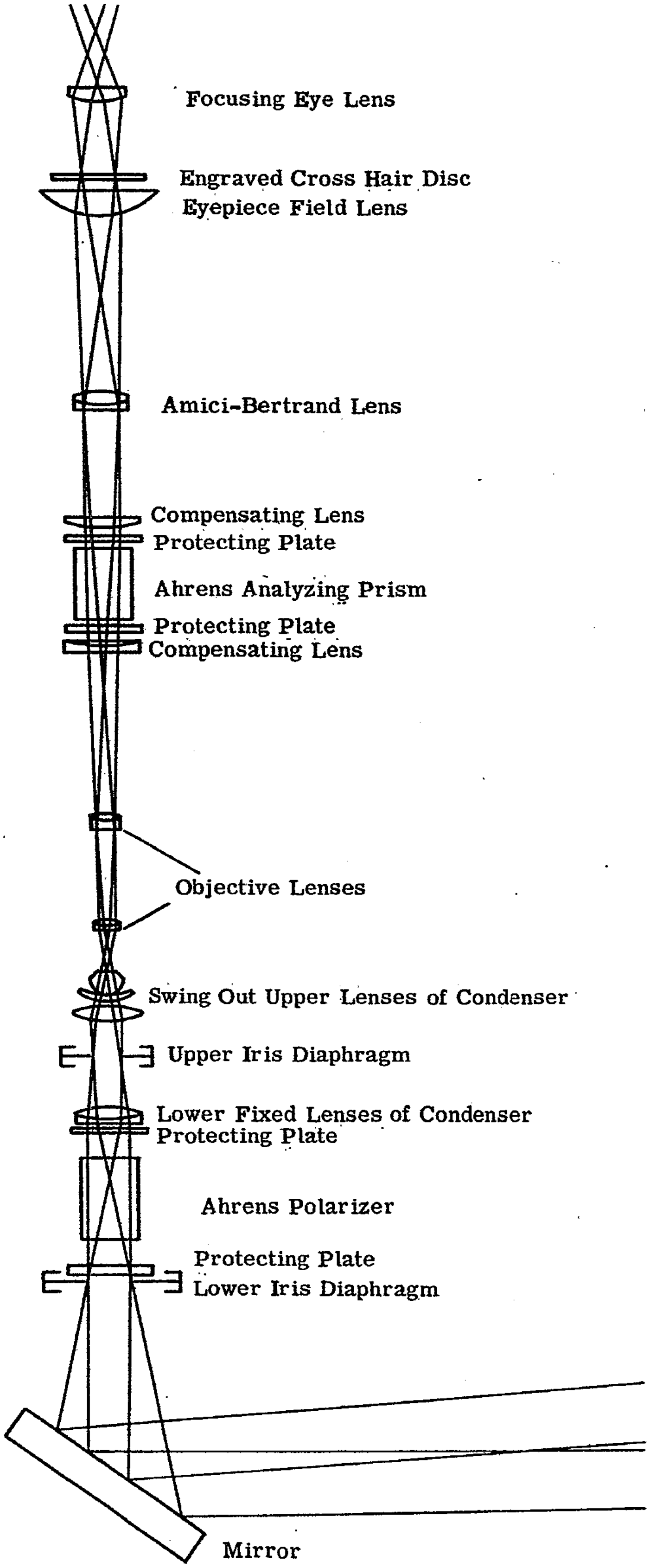


Figure 23. 22- Optical system of a petrographic microscope.

23.11.2 Characteristics.

23.11.2.1 Substage condenser. The substage condenser is made up of two parts one of which is designated as the lower fixed lenses of the condenser and the other as the swing out upper lenses of condenser. When these two systems are working together, the NA of the combination may be as great as 1.40. When only the lower part is used, the NA may be as low as 0.25. At the lower (front) focal plane of each of the above condensers, is located an iris diaphragm designated in Figure 23.22 as the lower iris diaphragm and the upper iris diaphragm. Beneath the lower condenser and between it and the lower iris diaphragm is the polarizer which may consist of a Nicol or Ahrens polarizing prism or a sheet of Polaroid. The polarizer is generally rotatable and provided with an angular scale. A detent stop may indicate the zero setting of the polarizer.

23.11.2.2 Objectives. The objective lenses used in the petrographic microscope are identical in design with the achromatic series of microscope objectives already mentioned. However these objectives must be free from strain otherwise their birefringence will interfere with measurements made upon mineral specimens.

23.11.2.3 Analyzing system. The analyzer may be a Polaroid plate or a polarizing prism of the Nicol or Ahrens type. The light passing from the objective to the eyepiece is convergent. Since a polarizing prism will produce astigmatism under such circumstances, it is necessary to parallelize the light traversing the polarizing prism. For this purpose a negative lens is used below the analyzer to cause the convergent light to become parallel. Above the analyzer is placed a convergent or positive lens of such focal length that the rays are focussed on the cross hairs of the eyepiece. These lenses need not be achromatic as the image forming bundles of rays are of such a small aperture. The introduction of these compensating lenses necessarily change the initial magnification and to avoid having different magnifications when the analyzer is inserted or withdrawn from its position on the axis of the instrument, the compensating lenses are fixed inside the body tube. The analyzing prism itself is protected against dust, fumes, and moisture by two windows labelled in Figure 23.22 as protecting plates. These plates should not be plane and parallel as there would be detrimental reflections between the surfaces of the plates. These windows should be in the form of menisci of zero power.

23.11.2.4 Amici-Bertrand Lens. This lens is located in such a position, and is of the correct focal length to image the back focal plane of the objective onto the cross hairs of the eyepiece. It will be seen that in this case the entire microscope becomes a telescope focussed for infinity. Of course, when the Amici-Bertrand lens is slid out of the instrument the system is a microscope. The Amici-Bertrand lens may be focusable and equipped with an iris diaphragm.

23.11.2.5 Eyepieces. The eyepieces in the illustration are of the Huygenian type with the eyelens focusable upon the cross lines of the reticle. The entire eyepiece is prevented from rotating by means of a tongue, or screw, in the eyepiece engaging a slot in the upper end of the body tube.

24 DESIGN PHASE OPTICAL TESTS

24.1 INTRODUCTION

24.1.1 Uses. Optical testing methods are widely used in all branches of scientific and technical work. The basic techniques, or modifications thereof, enable some of the most sensitive and precise measurements man has ever known. Gage blocks may be measured to better than 0.0000001" with relatively simple interferometric apparatus while velocities of satellites hundreds or thousands of miles away may be measured with the Doppler shift techniques common to older astrophysics problems.

24.1.2 Related fields. It will be noted that from time to time reference will be made to work that has been done in the field of microwave antennas. This has been done in the belief that it will be very instructive to become acquainted with design techniques involving wavelengths that are frequently approaching a tenth of the radiating aperture. Further, the use of aberrations, interference, diffraction, and control of aperture illumination are discussed and demonstrated in a way frequently difficult at light optics frequencies. The very recent achievements in light optics where the aberrations are all reduced (save color) to the diffraction-limited stage is many years old in the microwave-antenna field. Microwave antenna designers borrowed heavily from older optical techniques and it is quite possible that a study of their efforts will be highly rewarding to the light-optics designer.

24.1.3 Methods and problems pertinent to optics.

24.1.3.1 While these methods cover a wide gamut, discussion in this section will be confined to a small sampling of the methods particularly suitable to the design, construction, and evaluation of visual optical systems. A few words regarding the origin of the testing problems which will be encountered will be in order.

24.1.3.2 The design of an optical instrument is obviously predicated on a need having been established. Sometimes the nature of this need is such that electrical and mechanical considerations may dictate, to a considerable extent, the physical shape of the optical system. However, even after this has been determined there still remains the problem of translating the customer's purely optical requirements into a form that is significant to the lens designer. Field of view, curvature of field, transmission over a given spectral band, distortion, etc. can be specified rather accurately and unambiguously. Questions, however, as to image quality and what figure of merit is to be used in deciding whether this or that design will most closely give the customer the information he seeks when he uses the instrument, raise problems that have yet to be solved completely. There seems to be more and more evidence of late that to phrase the problem in this way--viz. that the optical instrument be an "information handling system" -- is preferable to the more vague requirement that it be a system that forms a good image. Agreed, the former actually sounds more vague, but current effort indicates the above sentence is probably correct.

24.1.3.3 The postulating of a figure of merit implies that one must test proposed designs to see if they meet the assumed theoretical criterion. Once the design is firm, the optician takes over and now he must perform tests to see that his construction faithfully follows the prescription given to him by the lens designer. Here we must point out that there is another testing step necessary. The optician's job may be considered complete and accurate when the radii, edge thicknesses, center thicknesses, spacings, indices, etc. agree with the specifications handed down by the lens designer. The fact that the optician's work is presumably accurate does not, however, serve as a complete check on the usability of the system. It must be remembered that the designer used some theoretical criterion such as amount of energy in a point image, phase front or Seidel aberrations. The next step therefore is to see how well the constructed system lives up to his predictions in one or more of these respects.

24.1.3.4 There is little doubt that the ultimate test of any system is a field test under the original conditions imposed in the customer's specifications. A system can conceivably be excellent in the laboratory and yet be so sensitive to vibration that it is useless in the field. Further laboratory testing under simulated field conditions is therefore indicated; installation in field equipment being attempted only after the prototype has been tested thoroughly in the laboratory.

24.1.3.5 Here is another point that should be strongly raised. Granted that field tests are the ultimate in one sense, we should not lose sight of the fact that the nature of field tests frequently is such as to cloud the performance of the optical system by the introduction of parameters not basically a part of the problem. The writer clearly recalls airborne cameras yielding several hundred lines per minute resolution in the laboratory and only 20-30 lines per minute in the air. The trouble was definitely not with the camera or optical system but rather with the mechanical mounting in the plane. Some more or less absolute standard of perfection based on the customer's optical requirements is therefore mandatory. Tests in this category are extremely valuable. Resolving power, sine-wave tests, etc. fall into this category.

24.1.4 The testing program.

24.1.4.1 A consideration of the principles outlined above indicates that the complete testing program rather naturally falls into the following categories. It should be pointed out that many more types of tests are known in each category, but space permits only this limited sampling.

24.1.4.2 Testing during the design phase.

(1) Calculation of the Seidel Aberrations

(2) Calculation of the Spot Diagrams

(3) Determination of the phase front and perhaps the predicted diffraction by knowing the phase front and amplitude distribution over the aperture.

24.1.4.3 Testing during the manufacturing phase.

(1) Foucault Test

(2) Star Test

(3) Ronchi Test

(4) Interferometric Tests and/or determination of phase front.

(5) Measurements of curvature of field, astigmatism, transmission, field of view, front and back focal lengths etc.

24.1.4.4 Testing during the evaluation phase.

(1) Any or all of the tests in 24.1.4.3 above.

(2) Measurement of the resolving power.

(3) Measurement of the sine-wave response.

We will now proceed to discuss each of these tests.

24.2 CALCULATION OF THE SEIDEL ABERRATIONS

24.2.1 Object-image relationship. From a strictly theoretical point of view, an optical system may be said to be perfect if its response is "collinear" i.e. points are imaged as points, lines are imaged as lines, and planes are imaged as planes. A further qualification is required--namely that the definition just given applies strictly and only to an optical system where the magnification is unity for all image points. While such systems do have significance, most optical systems require either minifications (telescopes, field cameras, etc.) or magnification (microscopes, etc.). We therefore qualify the concept of collinearity by adding that magnification or minification may exist, but should be constant for all points in the image. The above definition, even with its qualifications, applies more to photographic than to visual optical systems because of the reference to a flat focal surface. While curved focal surface systems have been used in photography, they are rare because of the practical problems involved in film handling. Almost all photographic systems require a flat focal surface, i.e. a focal plane. For visual optics we may relax this requirement somewhat. Indeed the ideal system is one whose curvature of field matches that of the eye.

24.2.2 The importance of Seidel Aberrations. It has been found possible by Seidel (1) to express the deviation of an actual image produced by a system, from the theoretically perfect system by a series expansion. This series expansion was given previously in Section 8. The monochromatic deviations from the ideal flat focal surface collinearity are called aberrations and include spherical aberration, coma, astigmatism, curvature of field and distortion. To the extent, then, that this series expansion accurately depicts what happens to an image point, the calculation of these Seidel aberrations constitute a powerful first approximation in the design of an optical system. It is equally clear that the calculation of these aberrations may be considered as a theoretical test of such a system. The method of calculating these aberrations, and the detailed significance of each has been previously treated. The subject is raised here again to point out the use of these aberrations in the theoretical tests which may be applied to an optical system. The reader should refer to Sections 8-10 for more details. It should also be pointed out that these aberrations are strictly geometrical and that two different systems may have the same aberrations and yet show quite different images due to the fact that the wave nature of light is completely ignored (except for the variation of index with wavelength).

---

(1) Seidel: Astronomische Nachrichten, 43, 289-332 (1856).

24.2.3 Seidel Tolerances. The criticism sometimes levied is that it is pointless to design a system on the basis of purely geometrical optics because of the neglect of interference etc. To our knowledge no optical system has been designed, at least in recent years, without reference to the wave nature of light. Frequently this is done by explicitly placing tolerances on the aberrations by reference to the Rayleigh (2) stipulation that the maximum path deviation from a given object point to a given image point be not more than $\lambda/4$. Discussions of this may be found in Conrady (3) and Martin (4). These optical tolerances are:

For primary marginal spherical,

$$\text{permissible primary } LA' = 4\lambda/N' \sin^2 U'_m \qquad (1)$$

For primary zonal spherical, (assuming $LA' = 0$)

$$\text{permissible } LZA' = 6\lambda/N' \sin^2 U'_m \qquad (2)$$

For primary Coma (Coma_s)

$$\text{permissible Coma}_s = \pm \lambda/2N' \sin U'_m \qquad (3)$$

For focal range,

$$\text{Focal range} = \lambda/N' \sin U'_m \qquad (4)$$

For astigmatism, Ast's

$$\text{permissible Ast's} = \lambda/4N' \sin^2 U'_m \qquad (5)$$

For curvature of field,

$$\text{permissible } X' = \text{focal range} = \lambda/N' \sin^2 U_m \qquad (6)$$

Note: U'_m is the angle between the ray and the axis, N' is the index of refraction in image space; and λ is the wavelength of the radiation.

24.2.4 Use of the Seidel Tolerances. One should use these tolerances with exceeding care particularly with high speed systems. This occurs because the focal range allowed by the $\lambda/4$ path difference criterion is assumed small compared with the actual focal length. Secondly the field angle is assumed sufficiently small so that $\sin^2 \frac{(U'_m)}{(2)} = 1/4 \sin^2 U'_m$. One should further regard these tests as representing a theoretical arbitrary standard which may be too tight or too loose in special circumstances. For fast systems (microwave antennas are a good example outside of the field of visual optics), the tolerance on spherical aberration as computed from (1) is too loose--usually by a factor of 4 or more. The tolerance on coma is too loose for many visual systems where the coma may be the most serious aberration and every attempt should be made to reduce it sensibly to zero. The astigmatic tolerance is usually too tight, and a lens may be expected to produce good results even if the astigmatic tolerance is exceeded by a factor of 2.

24.2.5 Conclusions. The subject of Seidel aberrations from purely geometrical optics is considered here in conjunction with tolerances imposed by physical optics because they have been the prime standards against which lenses were compared until the relatively recent present. Most lenses still are designed on this basis today although there are some who think that sine-wave response calculations may replace them in future years. In conclusion we may say that the reduction of aberrations to within, or at least close to, the stipulated tolerances is a necessary but not sufficient condition that to assure a lens so constructed will perform well. Actually the reduction of the aberrations to the specified limits results in a wavefront that is sensibly spherical in image space. The true image, however, involves amplitude as well as phase, and the Seidel aberrations give no explicit information regarding amplitude.

24.3 THE SPOT DIAGRAM

---

(2) Lord Rayleigh, Collected Papers, vol. 1, pp. 415-459.

(3) Conrady, Applied Optics and Optical Design, pp. 136, 395, 434 et seq., Dover, (1957).

(4) Martin, Technical Optics, vol. 1, p. 139, Pitman, (1948).

also Jacobs, Fundamentals of Optical Engineering, 443, McGraw Hill, (1943).

24.3.1 Introduction. In the past, the labor involved in doing any but the simplest of ray tracing was such that relatively few rays were traced in the actual lens design process. With the advent of electric desk calculators, it became possible to trace more rays in the same time. As a result tracing rays out of the meridional or tangential fan became more common. It was not until the relatively recent present that the designer was freed of this time limitation by the development of the high-speed, electronic-computing machinery. It is now possible to trace hundreds of rays in the same time it took to trace just a few some years ago. This has resulted in lenses being designed much more carefully than ever before. The aberrations determined by tracing rays as just discussed are definitely an approximation that is very good under some circumstances, but the usual Seidel third-order aberrations are frequently misleading: higher order aberrations sometimes being dominant.

24.3.2 Aspherics. Another factor brought into being recently is the use of aspheric surfaces. Desk calculators or no, tracing through aspheric surfaces can be a monumental task when done by hand. There is ample evidence, however, that freed from the restriction of purely spherical surfaces, the designer can almost always do a far better job with aspherics than he can with spherical surfaces.

24.3.3 Development and limitations. One of the first testing techniques that took full advantage of the power of the large computers was that evolved by Herzberger (5) and later by Hopkins (6) and was called the "spot diagram." In essence the entrance pupil is divided into equal areas, and a ray is traced through the center of each area--the assumption being that the energy represented by each ray is the same. The intersection of these rays with an assumed focal plane was a spot, hence the term "spot diagram." The more compact this spot, the more nearly perfect was the lens judged to be by the standards of geometrical optics. This is discussed in Section 8. We should thus clearly realize that this technique is restricted to non-diffraction limited systems. In this connection we should also realize that while most optical systems today are not diffraction-limited, there is a growing class of high precision systems widely emphasizing aspherics where the only aberration left is color, and where the performance is almost an order of magnitude better than it was ten years ago. For such systems, the spot diagram can serve only as a rough first approximation. The vast majority of visual and photographic optical systems are aberration-limited rather than diffraction-limited so the spot diagram is still a powerful tool.

24.3.4 Techniques. There are basically two techniques for getting a spot diagram. In one the required number of rays is actually traced, and the intersection points with the assumed focal surface are plotted. In the other a relatively small number of rays is plotted, and the intersection coordinates of the others are obtained by an interpolation and extrapolation process developed by Herzberger. It should be noted here that the interpolation process does more than just give the intersection points. Via the series expansion required for the interpolation it also gives a set of terms not unlike those of Seidel. The difference is major, however, in that the Seidel aberrations work particularly well near the axis while the "Herzberger aberrations" fit well over the entire aperture. Space does not permit us to go more deeply into this use of spot diagrams, but the reader is encouraged to refer to Herzberger's articles on this subject (5), (7), (8).

24.3.5 Examples. Those interested in this subject are also urged to obtain National Bureau of Standards Report No. 5640 entitled "Numerical Analysis of a 6" f/3.5 Aerial Camera Lens (006BC035 - 15)". This report by Stavroudis and Sutton shows clearly the extent to which the spot diagram testing is currently employed. Not only are the spot diagram shown for various assumed focal plane positions and angles of obliquity, but also the values of vignetting, distortion, chromatic aberration, energy distribution, and resolving power are derived for this lens directly from the spot diagrams. It is interesting to note the excellent correction that seems to have been achieved in this lens. For full aperture the diameter of the Airy disk is 4.0 microns. If we inspect the following table, Table 24.1, taken from Stavoroudis report, we see that 80% of the total points fell within a circle on axis whose diameter was 3.93 microns. For an aberrationless system theory indicates there will be 83% of the total energy within the Airy disk. The close agreement between theory and spot diagram prediction indicates the excellence of the design, at least for on axis work. In another series of experiments Stavroudis and his colleagues at the National Bureau of Standards calculated the spot diagram of a completed lens. The comparison of the spot diagrams and corresponding actual photographs for two given positions is shown in Table 24.1.

24.4 PHASE FRONT CALCULATIONS

24.4.1 The spherical wavefront. It has been pointed out that the Seidel Aberrations, when they are fully corrected, result in a spherical wavefront converging on the image point. Modern computing machinery has enabled the designer to calculate directly the wavefront and thus determine not only the phase errors over the aperture but where the focal point should be placed.

(5) Herzberger, J. Opt. Sec. Am 37, 485 (1947).
(6) Hopkins, J. Opt. Sec. Am 44, No. 9, 692-698 (1954).
(7) Strong, Concepts of Classical Optics, Appendix L by Herzberger, p. 537, Freeman (1958).
(8) Herzberger, Optical Image Evaluation, National Bureau of Standards Circular No. 526, U. S. Gov't. Printing Office (1954).

% Total points	0° µ	7° µ	11° µ	14° µ
10	0.674	3.66	4.88	4.02
20	1.22	6.54	12.2	13.2
30	1.69	14.0	25.6	27.2
40	1.97	24.7	43.7	45.7
50	2.43	37.8	66.2	67.6
60	3.05	53.3	89.3	92.2
70	3.71	71.7	118.	119.
80	3.93	97.8	149.	148.
90	6.08	132.	187.	189.
100	12.1	247.	266.	311.

Focal length = 5.972460

Plane of best focus at -0.042 mm

Table 24.1- Energy Distribution 006 BC01515.

The table gives the diameters of the smallest circles containing specified percentages of the total number of points in each of the four spot diagrams at the plane of best focus. The common center of the circles for a given spot diagram was taken where the density of the points appeared greatest.

The diameters are listed in microns to three significant figures. Note that the diameter of the Airy disk for a perfect lens as a full aperture of f/3.5 is 4.0µ. Dr. R. N. Wolfe of Eastman Kodak Co. Research Laboratories made a similar series of experiments in 1947 in conjunction with some of Herzberger's early work in this field (9). The subject has been extensively investigated as regards automatic data reduction by Goetz and Woodland (10) at IBM. Miyamote (11), Keim and Kapany (12) as well as many others have studied this very interesting optical test.

24.4.2 Technique. There are many ray tracing programs that will give this information. The one developed by Feder(13) is offered here. Again the techniques of using this method of testing are varied but the following one is typical. See Figure 24.2. Three or more rays are traced from plane PP through the entrance pupil, the optical system into image space. The entrance pupil is EE. Frequently among the rays of interest are the upper rim ray (U), principal ray (Pr), and lower rim ray (L). A point B' on the principal ray in image space is picked arbitrarily and, from the ray tracing data, the optical path length BB' is determined. From the ray tracing data for rays U and L as well as those originating at other points (frequently zonal) such as D and F, optical path lengths equal to BB' are laid off along the rays. The termination points C', D', F' and K' are then marked and the curve passing through them constitutes the equiphase front in the plane of the paper. The deviations from a perfect circle (or sphere in three dimensions) are clear and corrections may be made as necessary.

24.4.3 Applications and limitations. This phase front technique has long been used in the design of microwave antennas because of the optical simplicity (generally speaking) of such systems. It is particularly useful in optical design as the phase front may be determined experimentally by long established techniques. This gives the designer an immediate check on how well the optician has fulfilled the prescription given to him. It should be pointed out that the diffraction pattern may now be determined, providing the amplitude distribution over the front is known. In some cases it is simpler to use basically the same technique but actually determine the phase variation over the exit pupil. The amplitude distribution over the exit pupil is determined and the diffraction pattern calculated as before. The possibility of varying the amplitude over the aperture by control of aperture shape, variation of transmission, or illumination with radius has been known for some years. A few of the efforts in this direction are the work of Conder and Jacquinot (14) in spectroscopy, the work of Osterberg and Wilkins (15) with microscope objectives, and the work of Silver (16) on tapered illumination of microwave antennas.

---

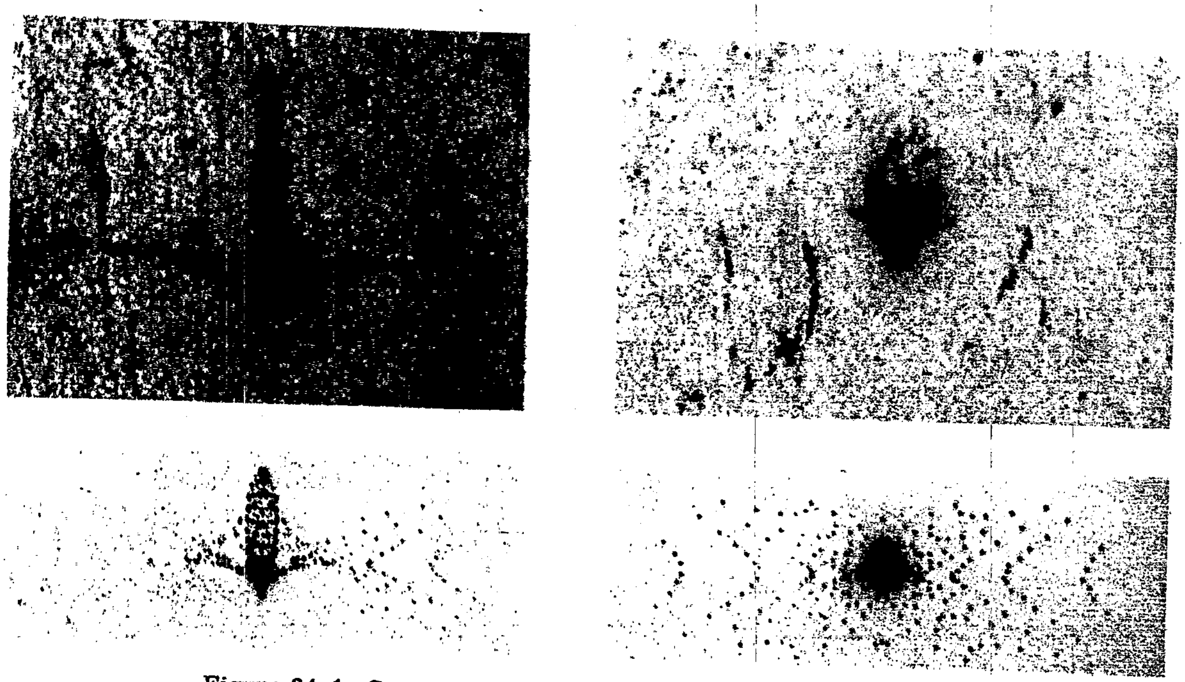
(9) Herzberger, J. Opt. Sec. of Am. 37, 485 (1947).

(10) Goetz and Woodland, J. Opt. Soc. of Am. 48, 965 (1958).

(11) Miyamoto, J. Opt. Soc. of Am. 48, 57, (1958); 48, 567 (1958), and 49, 35 (1959).

(12) Keim and Kapany, J. Opt. Soc. of Am. 48, 351 (1958).

(13) Feder, J. Opt. Soc. of Am. 41, 630 (1951).



20° Off-axis at Gaussian Focus

10° Off-axis, 0.3mm in from Gaussian focus.

Figure 24.1- Comparison of spot diagram and actual photograph.

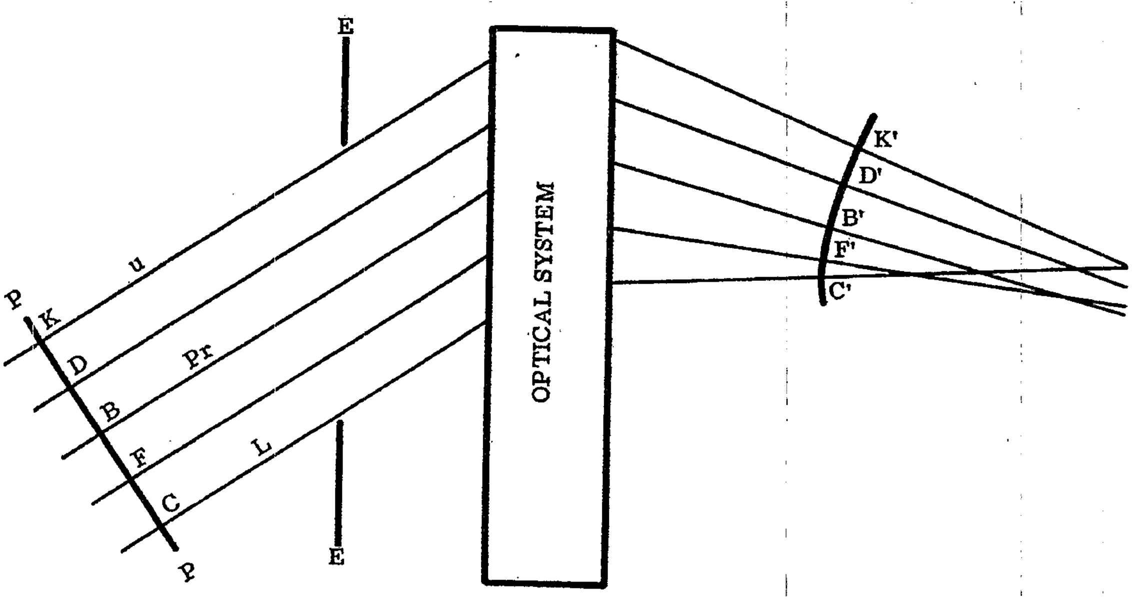


Figure 24.2- Determination of constant phase front from ray tracing data.

---

(14) Conder and Jacquinot, "Méthode pour l'observation des radiations de faible intensite au voisinage d'une raie brillante" Compte Rendus de l'Académie des Sciences (Paris), 208, 1639, (1939).

(15) Osterberg and Wilkins, "The Resolving Power of a Coated Objective," J. Opt. Soc. of Am. 39, 553 (1949).

(16) Silver, "Microwave Antenna Theory and Design" 187, McGraw Hill (1949).

25 PRODUCTION PHASE OPTICAL TESTS

25.1 INTRODUCTION

25.1.1 General. Intrinsic in the design of most optical systems is the calculation of the Seidel aberrations. In this section we will outline procedures for measuring these aberrations experimentally during either the production or the evaluation phase. In this connection it should be noted that over the years different laboratories have developed their own techniques for making these measurements. Frequently the difference between techniques is not so much a matter of difference in basic principle, as it is in the equipment that a particular laboratory happens to have on hand. While there are, then, many, many different ways to make each measurement, we will limit ourselves to one example of each. The interested reader may consult the references for additional information.

25.1.2 Theory vs practice. Before leaving this introduction to the measurement of Seidel aberrations, a few words of caution are in order. Aberrations may be completely isolated only in theory. The actual image embodies, simultaneously, all aberrations pertaining to it. This, of necessity, complicates the measurement, and particularly complicates the detailed checking of the theoretical predictions as to the values of the individual aberrations. It should be pointed out also that the accuracy with which the aberrations need to be measured is a function of the importance of the particular aberration to the job at hand. The experiments to be described generally assume a white light source. Chromatic effects are determined by use of the appropriate filters.

25.2 FOCAL LENGTH

25.2.1 Importance of focal length. Certainly one of the fundamental constants of any optical system that is of prime importance in the evaluation of the significance of all Seidel aberrations is focal length. Not only is the value of the focal length of importance, but also a precise statement of the point from which the focal length is to be measured is mandatory. Some years ago an aerial camera lens was designed and simultaneously the camera body was fabricated, presumably for the same focal length systems. When the lens was installed in the body, a photographic check showed hardly any semblance of an image. To say that it was "out of focus" was charitable. The error was tracked down ultimately to the fact that focal length meant measurement from the secondary principal point to the lens designer and meant from the rear surface to the machinist. Through human error, both lens and body were allowed to be fabricated on this erroneous basis. Since the secondary principal point lay several inches inside the lens, the horribly blurred image was not surprising. Recently a similar situation developed in a missile-tracking system where the optical designer measured the focal point with respect to the front surface and the machinists assumed that it was measured from the rear surface (the optics involved a thick mirror). Since the system was quite fast with short focal length, the one inch central thickness of the thick mirror played havoc with the performance of the system when finally assembled.

25.2.2 Measurement of focal length. While there are many methods for measuring the focal length of an optical system (1) (2) (3) (4), one of the most accurate for lenses of medium focal length employs the nodal slide. A photograph of one in use at the National Bureau of Standards is shown in Figure 25.1. The essential part of the nodal slide is the provision for moving the lens system longitudinally with respect to a vertical axis of rotation. This vertical axis is mounted so that it may be positioned longitudinally with respect to a collimator of appropriate size. Usually the object for the collimator is a very small point source set at the focal point of the collimator.

25.2.3 Test setup. The equipment is set up as shown schematically in Figure 25.2. In use the magnifier or microscope is set up approximately at the focal point. The lens under test is then moved backward and forward along the nodal slide until rotation of the nodal slide through a small angle, B, produces no sidewise shift in the image. The focal length is then the distance between the axis of rotation of the nodal slide and the appropriate focal point of the magnifier or microscope. There are many variations on this technique, some employing auto collimation, some focusing the image on a card, etc. Negative optics may also be tested in this manner by the addition of a positive lens of known characteristics. Knowing the position of the secondary nodal point (coincides with the principal point if the index of refraction of image space is air), the focal length may be specified with respect to the vertex of the rear surface (this distance is known as the "back focal length") or to any other convenient part of the lens.

---

(1) Cheshire, Trans. Optical for (London) 22, 29 (1920-1921).
(2) Kurtz, Jour. Opt. Sci. of Am. and Rev. of Sci. Instr. 7,103 (1923).
(3) Searle, Experimental Optics, Exp. 37, 185 Cambridge Univ. Press (1925).
(4) Wagner, Experimental Optics, Exp. 67, 136, Wiley (1929).

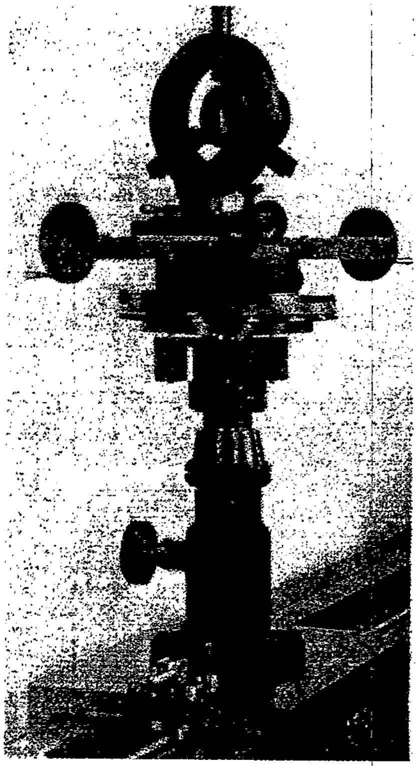


Figure 25. 1- Nodal slide developed at the U. S. National Bureau of Standards.

L_c = collimating system

L_x = system under test

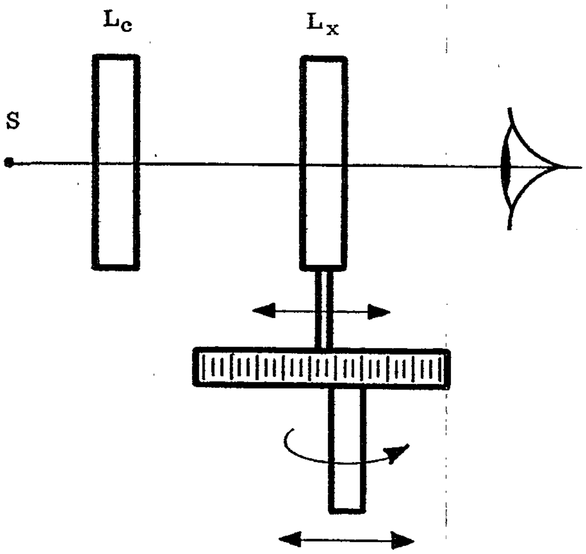


Figure 25. 2- Measurement of focal length by use of visual nodal slide.

25.3 LONGITUDINAL SPHERICAL ABERRATION

25.3.1 On-axis performance. Of considerable importance in almost all optical systems is the on-axis performance. Since the principal use of many visual optical systems is tracking in one form or another, systems are ultimately pointed directly at the target. The nature of the image on-axis is thus important. A factor also of much significance is the degree to which the system may be "opened up" and still maintain a good image. This latter requirement involves spherical aberration.

25.3.2 Hartmann test. While there are many techniques for doing this (5) (6), the simplest is perhaps the Hartmann test (7) (8) (9). It may be done either photographically or visually and can be made reasonably sensitive. It is probably not as accurate as a newer method developed by Washer, but is chosen here for its directness and simplicity.

25.3.3 Test procedure. Blocking off all but holes 1 and 8 in the Hartman Disk shown in Figure 25.3 will give the marginal focus. Holes 2 and 7 should be located at the zonal positions, and their intersection will give the zonal focus. The paraxial focus may be determined with holes 4 and 5, or by direct inspection of the lens stopped down to a very small circular aperture. From the data just obtained the longitudinal spherical aberration may be determined. Filters may be used to get the aberration for different colors if so desired.

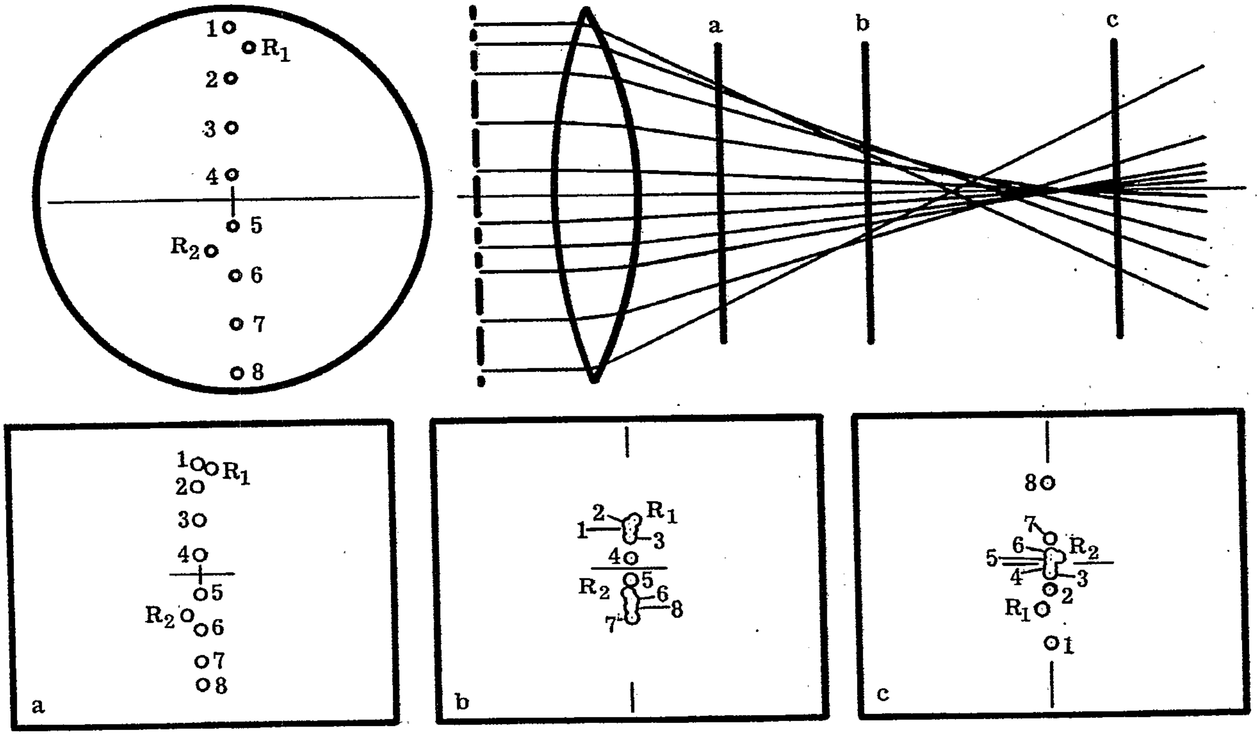


Figure 25.3-Measurement of spherical aberration by the Hartman Test.

(After Strong's, Concepts of Classical Optics, W.H. Freeman and Co. 1958)

---

(5) Washer, Jour. of Res. of Nat'l Bureau of St'nds 61, No. 1, 31, (July 1958).
(6) Monk, Light-Principles and Experiments 349, McGraw-Hill (1937).
(7) Strong, Concepts of Classical Optics, 354 Freeman (1958).
(8) Hartmann, Zeit. f. Inst. XX IV, 1 (1904); and subsequent papers in 1904.
(9) Bureau of Standards Scientific Papers No. 311 and 494.

25.4 COMA

25.4.1 Asymmetrical flare. Asymmetrical flare produced by coma is one of the most important aberrations to eliminate. The reason for this is that most of the other aberrations produce an image degradation that is more or less symmetrical with respect to the principal ray. For example this means that even though astigmatism may be present, the system may be pointed with a high degree of accuracy by centering the point of greatest density of the image on the cross hairs, etc. When the image is degraded asymmetrically, this same procedure can produce a pointing, or boresight, error.

25.4.2 Collimator check for coma. Coma, being an off-axis aberration, is somewhat difficult to separate from astigmatism. Usually in testing optical systems the optician will simply use a collimator to illuminate his lens at successive angles off axis. The focal plane image is then studied, and, if the flare is more than that allowed in the specifications, the system is reworked.

25.4.3 Hartmann disk. Coma may be demonstrated to a fairly successful degree by use of the Hartmann disk placed before the lens with the lens illuminated by off-axis parallel light and image space then studied, as indicated in the measurement of spherical aberration. Another simple method for measurement of coma using the Hartmann method is described in Hardy and Perrin (10), and refers to a method described previously in the National Bureau of Standards Scientific Papers No's. 311 and 494.

25.5 ASTIGMATISM AND CURVATURE OF FIELD

25.5.1 Measurement of astigmatism. Astigmatism may be measured accurately by a series of Foucault Tests with the knife edges at right angles in the basic manner described in the section devoted to the Foucault test. It may also be measured quite simply by the arrangement illustrated in Figure 25.4. In practice, the lens under test, L_x, is rotated and the traveling microscope, M, is adjusted until the image of the reticle, R, is found. The microscope is then adjusted; first until the vertical lines are in best focus; then until the horizontal lines are in best focus. L_x is then rotated to successive angles up to the maximum field angle, and the positions of best focus as just described measured at each angle. A plot of the positions of best focus for the vertical lines will give the sagittal (secondary) focal surface, and the corresponding plot of the positions of best focus for the horizontal lines will give the tangential (meridional or primary) focal surface. The reason why the tangential plane images the horizontal lines best is shown clearly in Jenkins and White (11). The microscope must of course be movable laterally as well as longitudinally to examine the astigmatism at various points on the focal plane. If in addition to making determinations of the position of best focus of the horizontal and vertical lines, one also notes the position of best overall focus of the central point in the grid system, the curvature of field may be determined.

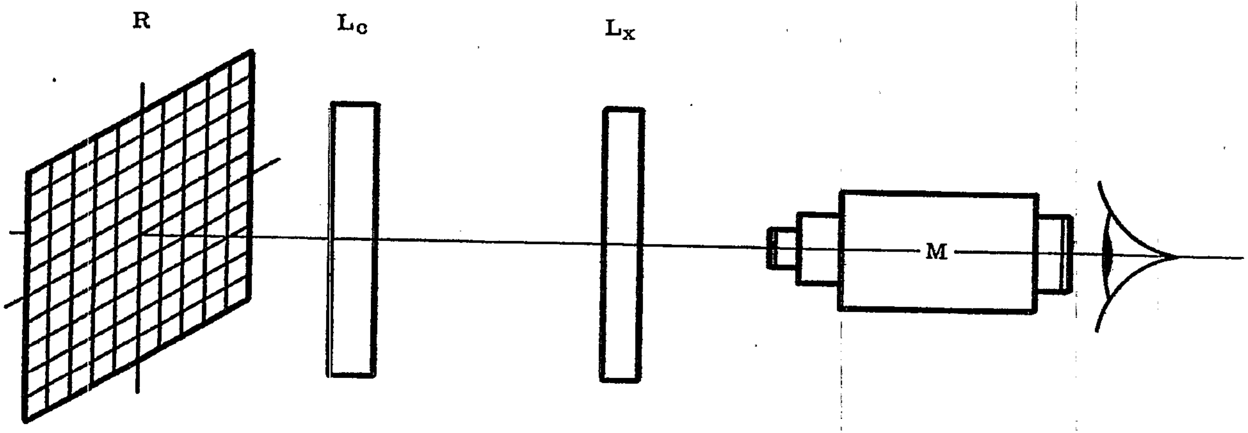


Figure 25.4- Measurement of astigmatism with a grid reticle.

---

(10) Hardy and Perrin, The Principles of Optics, 382, McGraw-Hill, (1932).

(11) Jenkins and White, Fundamentals of Optics, (2nd edition), 139, McGraw-Hill, (1950).

25.5.2 Determination of curvature of field. Generally, curvature of field may be determined with respect to the flat focal plane of a camera by placing a flat glass plate with a grease pencil mark across a diameter in the position occupied by the film. The grease pencil mark is towards the camera lens. One then notes the position of the traveling microscope when focused on the grease pencil mark. The camera being set up similar to that in Figure 25.4, the microscope is then focused on a star image (or central point of the reticle). The difference between the two readings is a measure of the curvature of field. Curvature of field measurements are very important in visual systems as the curvature of field of the object must match that of the eyepiece or else considerable image degradation ensue.

25.5.3 Consistency of test procedure. One should note clearly in all of these testing methods that if a system is to be used visually, then ideally it should be tested visually. Photographic testing does have its advantages, however, as it furnishes a record.

25.6 DISTORTION

25.6.1 Importance of distortion study. In optical systems designed essentially for visual observation and study of objects on-axis, the aberration known as distortion is really not too important. There are many systems, however, where, while the target may be centered in the eyepiece, measurements must be made over the entire field of view. An example of such a system is a rangefinder.

25.6.2 A rapid check for distortion: Distortion may be measured photographically very simply by replacing the microscope in Figure 25.4 by a good quality camera, known to be well corrected over the field of the optical system under test. The grid reticle is then photographed and the distortion, whether pin cushion, barrel, or irregular, is immediately obvious when D is set = 0.

25.6.3 An accurate distortion measurement for small optical systems. An excellent method of making this measurement for small optical systems with the basic nodal slide has been outlined by Washer, Tayman, and Darling (12). The procedure is as follows:

(a) The optical system under test is placed on the nodal slide shown in Figure 25.2.

(b) A measuring microscope is adjusted with respect to the lens until a focus is found.

(c) The lens system is then moved in the usual way along the nodal slide until small rotations (the microscope having been kept in focus by longitudinal movement) show no lateral movement of the image.

(d) Assuming that the focal length, f, is now measured or known, it is clear that if the microscope were moved off-axis yet remaining in this focal plane that the distance to the lens would now be f sec β.

(e) If now, instead of moving the microscope, the lens is rotated in the nodal slide by an angle β and moved longitudinally a distance (f sec β - f) or f (sec β - 1) toward the collimator, the microscope should again see the image clearly. Actually the image will probably not be on-axis and the microscope will have to be moved laterally a small distance to pick it up again.

(f) The reading of the micrometer measuring this lateral shift is noted as R.

(g) The lens is now rotated through an angle $-\beta$ and the microscope, when repositioned, gives a reading L.

(h) The distortion, D, at the angle β, is then given by the simple expression,

$$D_\beta = \frac{(R-L)}{2} \sec\beta \qquad (1)$$

25.7 AUXILIARY OPTICAL MEASUREMENTS

25.7.1 Introduction. In the fabrication and testing of optical instruments, it is frequently necessary to make measurements that are made considerably less frequently in regular machine shops. One of these measurements is the radius of curvature of spherical and aspherical surfaces: another is the measurement of the index refraction.

25.7.2 Radii of curvature.

25.7.2.1 The radius of curvature of an optical surface whose diameter is on the order of 1 - 3" may be done very conveniently with a spherometer (13). This device takes many forms--one of which is shown in Figure 11.9.

(12) Washer, Tayman, and Darling, Journal of Res. of N.B.S., 61, No. 6, 509 (1958).

(13) op. cit., (10), pg. 366.

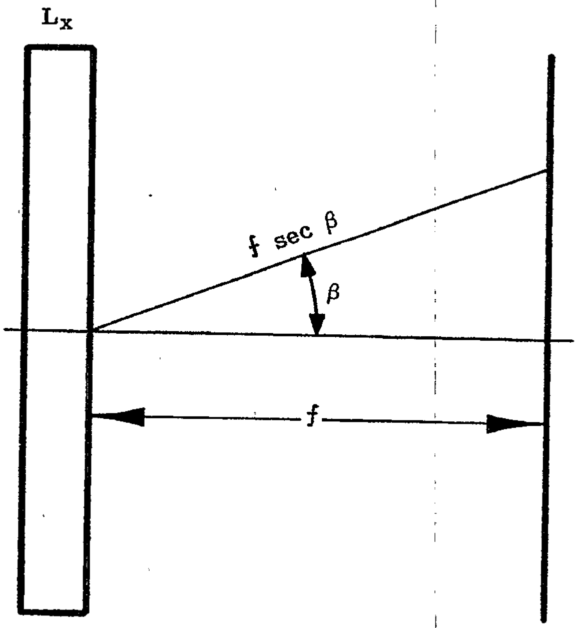


Figure 25.5- Basic diagram for measurement of distortion by nodal slide.

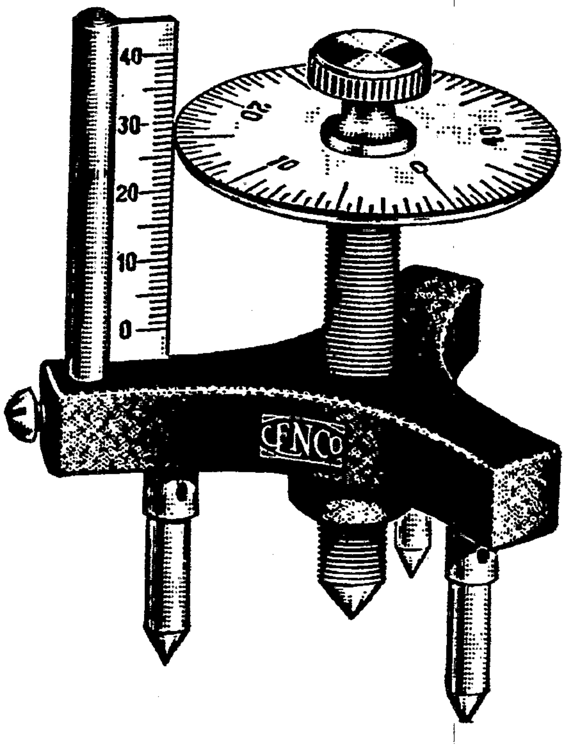


Figure 25.6- Elementary spherometer.

25.7.2.2 In use, the central spindle is screwed up and the three legs placed on the surface in question. The spindle is screwed down until it just touches the surface. The spherometer is then placed on a flat surface and the distance, S, the spindle must be advanced to meet the flat is noted. The procedure is reversed for a concave surface. The radius of curvature of the surface is then obtained from the following equation,

$$r = \frac{d^2}{6s} + \frac{s}{2} \qquad (2)$$

where d = the average distance between the legs.

25.7.2.3 For large surfaces the Foucault method described in paragraph 25.8.2 is used. For very small surfaces, less than about an inch across, a different method is employed. A provision is made for illuminating from the side, the cross hairs or reticle of a Gauss eyepiece, or equivalent, in a microscope or short focal length telescope (the choice depending upon the curvature of the sample to be tested). The microscope is focussed first on the surface of the sample, and the longitudinal position of the microscope recorded. The microscope is then racked back until there is no parallax between the illuminated cross hairs and the image from the surface. The cross hairs are then at the center of curvature. This position of the microscope is also recorded. The difference between the two positions is the radius of curvature. A telescope would be used in exactly the same way for greater radii of curvature. A similar technique can be used for positive surfaces.

25.7.3 Index of refraction.

25.7.3.1 Where it is possible to grind and polish a small sample of the material, the spectrometer furnishes a very fundamental method for measuring the index of refraction. The theory and method are outlined in Hardy and Perrin (14) and Sawyer (15). With a good spectrometer the values of the index so determined are good to ± .00003. A high precision spectrometer is shown in Figure 25.7.

25.7.3.2 For many samples it is not possible to get the sample in the form required for the spectrometer and for these the refractometer is frequently well suited. One of the many refractometers is the Pulfrich (16). This method, based on refraction at the critical angle, will give values correct to ± 2 parts in the fifth decimal place, and is shown schematically in Figure 25.8.

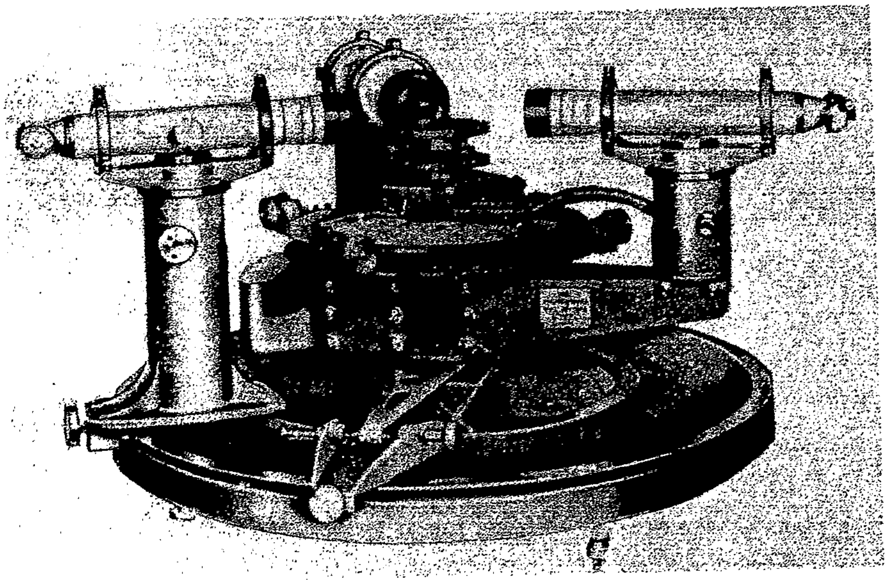


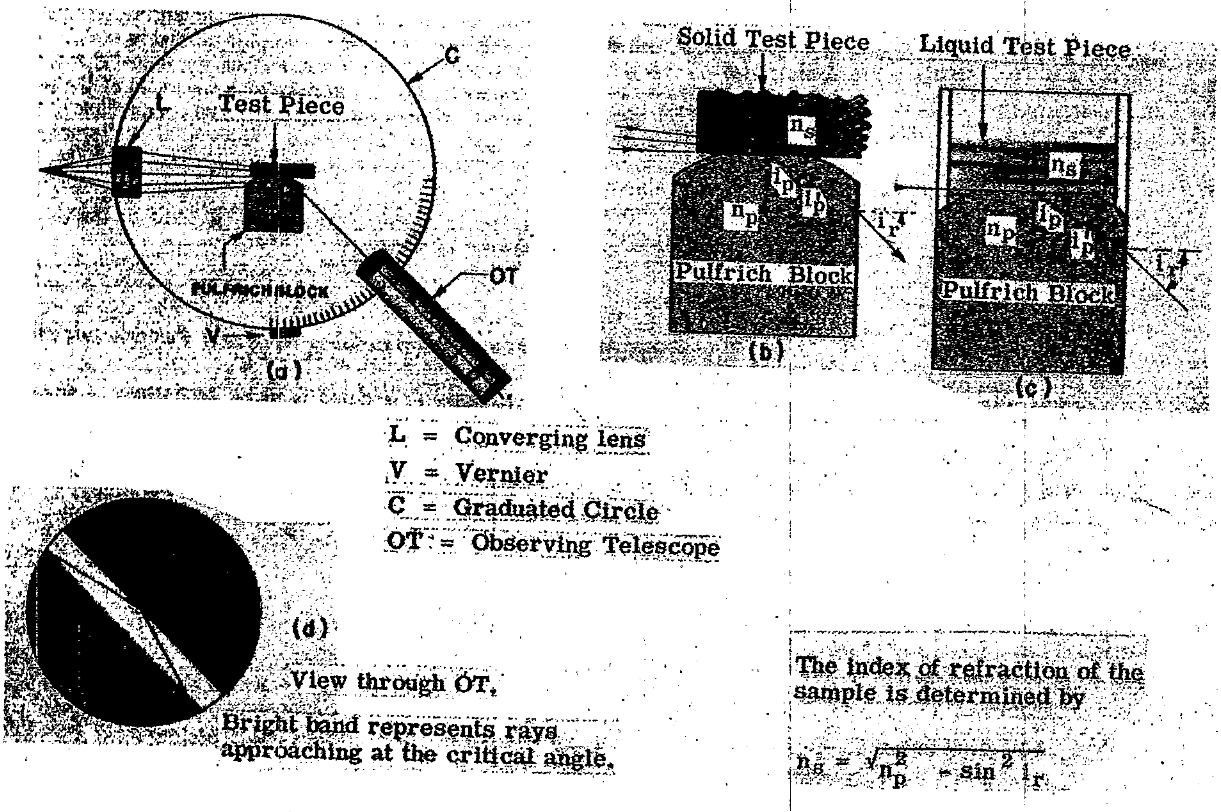
Figure 25.7 - The Guild-Watts precision spectrometer.

---

(14) op. cit., (10), 549

(15) Sawyer, Experimental Spectroscope, 55, Prentice Hall, (1944).

(16) op. cit., (10), 350.



L = Converging lens
V = Vernier
C = Graduated Circle
OT = Observing Telescope

Bright band represents rays approaching at the critical angle.

The index of refraction of the sample is determined by

$$n_s = \sqrt{n_p^2 - \sin^2 i_r}$$

Figure 25.8 - Schematic of the Pulfrich refractometer.

25.8 OPTICAL DEVICES, TESTING SYSTEMS AND PROCEDURES

25.8.1 Interferometry principles.

25.8.1.1 The most common and simplest method for testing flatness of polished surfaces of glass or other transparent material utilizes interference fringes that are formed between the tested surface S_1 and an optically flat surface S_2 as illustrated in Figures 25.9 and 25.10. The preferred source of light is an unfiltered, tubular, Cooper-Hewitt lamp, L, which is provided with a diffusing reflector, R, and a diffusing glass plate, G. The advantage of the arrangement of Figure 25.9 is that it permits the interference fringes to be viewed at normal incidence. The positions of the light source and the eye may be interchanged. Figure 25.10 illustrates the most common arrangement where viewing at normal incidence is sacrificed to gain greater freedom as regards working space. The light emitted by the Cooper-Hewitt mercury lamp is preponderantly green. It can be rendered quite monochromatic at 5461 A, whenever desired, by means of readily available optical filters that can be held near the eye. The interference fringes are usually viewed in unfiltered light. Contrast in the fringes is improved by placing black felt or paper beneath the optical flat in the manner indicated.

25.8.1.2 It should be noted that the interferometer surfaces S_1 and S_2, Figures 25.9 and 25.10, are in close contact. Excess dust and other dirt must be removed in order to reduce the thickness d of the air-film between surfaces S_1 and S_2. Because the separation d of the surfaces S_1 and S_2 is small, the resulting interference fringes belong to a select class of fringes known as Fizeau or as Newton's fringes. The principles underlying these fringes have been discussed in paragraphs 16.12.1.2 and 16.13.1.5. These fringes are characterized by the following important properties and propositions.

(a) The interference fringes appear in good contrast when one focuses upon the air-film between the interferometer surfaces, S_1 and S_2, provided that the reflectances of these surfaces are approximately alike. It is often said that the fringes appear to be localized in the film.

(b) Because the separation, d, of the interferometer surfaces is small, the location laterally of the fringes between the surfaces does not depend markedly upon the angle of incidence, provided that one views the fringes approximately along the normal to the surfaces S_1 and S_2.

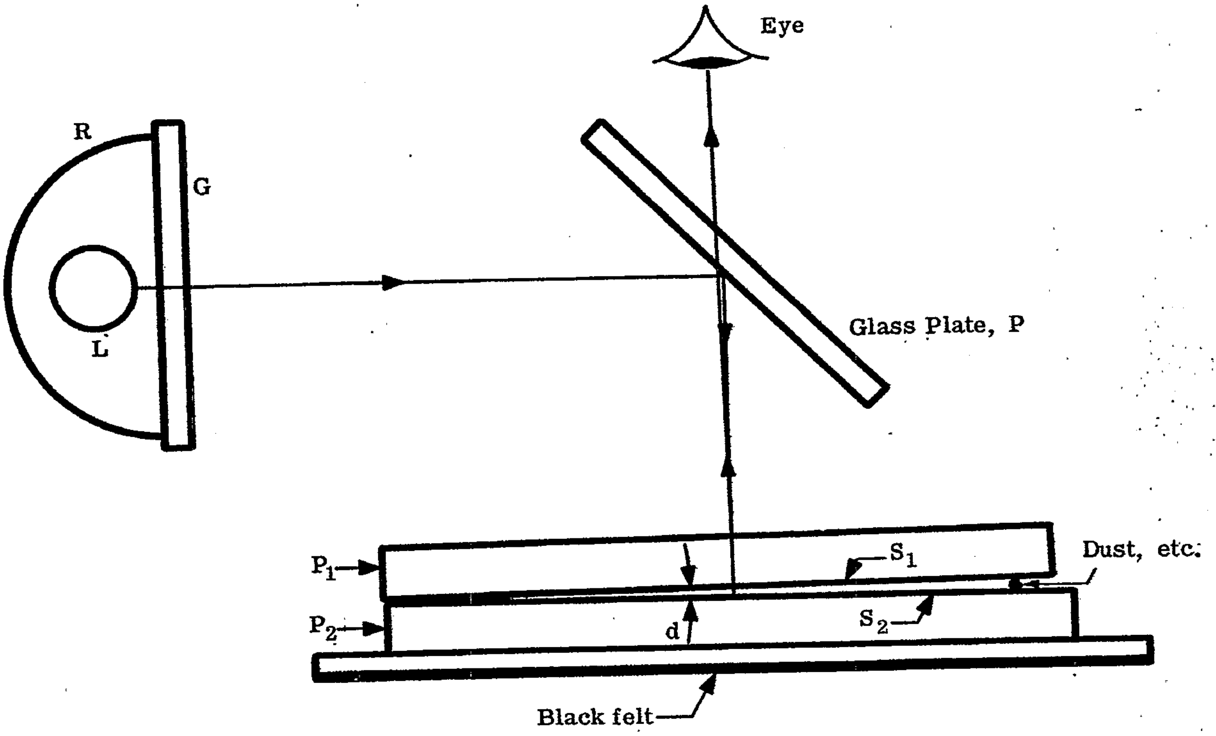


Figure 25.9 - Simple interferometer for viewing Fizeau fringes or Newton's fringes at normal incidence.

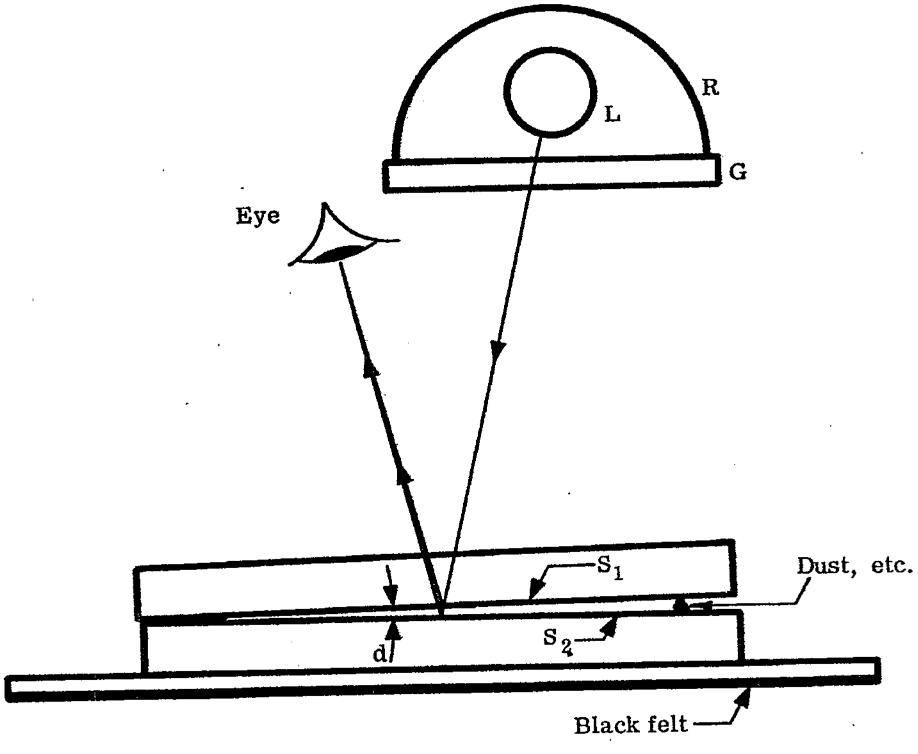


Figure 25.10- Most commonly used method for illuminating the interferometer. The fringes formed between surfaces S_1 and S_2 are observed at near - normal incidence.

(c) Each fringe marks the locus (lateral) of points for which the separation, d, is a particular constant. This constant is different for each fringe.

(d) When either surface is moved or distorted by the application of force or heat, each fringe moves in such a direction as to maintain the constant separation, d, associated with that fringe.

(e) Upon passing from one fringe to the next fringe of equal brightness or darkness, the separation, d, changes by one half wavelength.

(f) Upon passing from a bright fringe to the next dark fringe, the separation, d, changes by one fourth wavelength. It is assumed tacitly in (e) and (f) that the surfaces do not possess discontinuous jumps.

(g) When surfaces S_1 and S_2 are of nonabsorbing materials such as glass, dark fringes occur at separations, d, for which

$$d = \nu \frac{\lambda}{2} \; ; \; \nu = 0, 1, 2, 3, 4, \text{etc.} \tag{3}$$

and bright fringes occur at separations, d, for which

$$d = \mu \frac{\lambda}{4} \; ; \; \mu = 1, 3, 5, 7, \text{etc.} \tag{4}$$

wherein λ denotes wavelength.

Of these propositions and properties, (c) and (d) are of the greatest importance to the maker of optical flats. These two propositions or rules enable him to interpret the fringes for high and low areas on the surface under test. The optical worker recognizes fringes as contour lines on a contour map of his surface. Movement of the fringes upon application of pressure serves to distinguish the up-hill direction. Propositions (e) through (g) reveal the heights between contour lines. It is emphasized that propositions (e) and (f) are more general (and hence more often correct) than (g).

25.8.1.3 As one tests for flatness of a surface, the fringes become straighter as the surface becomes flatter. The effect of reducing the angle of the air-wedge between surfaces S_1 and S_2 by the removal or crushing of dust particles, is to widen the fringes and to increase their curvature except when S_1 is optically flat. Let, h denote the fringe width, i.e. the distance from one bright fringe to the next. For optical flats of high quality, the degree of flatness can be specified by requiring that any fringe shift from straightness shall not exceed a stated fraction of the fringe width, h, over a stated diameter or other dimension of the tested surface. As an example of the sensitivity of the method, a fringe shift h/5 corresponds to a change of separation, d, by the amount $\lambda/10$. Fringe shifts smaller than h/10 become difficult to detect and to measure in this type of interferometer.

25.8.1.4 In a second, and often preferred test for optical flats of high quality, the surfaces S_1 and S_2 are placed so closely in contact or are rendered so nearly parallel that a single fringe spreads over surface S_1. This broad fringe is examined for uniformity of color with an unfiltered source of light. It is customary to specify without further qualification that the surface shall be "polished to a uniform color".

25.8.1.5 The following procedure applies to that great class of test cases in which the departure of more than one fringe from flatness is tolerated. If the test surface is convex, only one area will contact the reference flat and this area will be surrounded by a number of alternately dark and bright Newton's fringes. The specification of flatness may be stated as the number of allowable Newton's fringes per inch, or other unit. If the test surface is concave, a ring-shaped area will contact the reference flat. The number of concentric fringes within this area can be counted and compared with the maximum tolerable number of Newton's fringes per inch or other unit. In practice, the surface S_1 is likely to display one or more convex or concave areas. Close examination of the fringes will distinguish between these convex and concave areas. If pressure is applied to a convex area in such a manner as to reduce the separation, d, between surfaces a given fringe about the area of closest contact must move outward from its center in order to maintain the locus of points for which d is a given constant.

25.8.1.6 No essential modification of the method of Figures 25.9 or 25.10 is needed for testing spherical surfaces. The reference flat, S_2, is replaced by a concave or convex reference surface whose radius is equal to the desired radius of the completed test surface. Suppose that surface S_1 has a smaller radius than surface S_2 as illustrated in Figure 25.11. Concentric Newton's fringes will appear around point O. The maximum allowed number of Newton's fringes per inch along the radial direction from O may be stated as the permissible departure of surfaces S_1 from the "test glass" having the surface of reference S_2. For surfaces, S_1, of high optical quality, the radius of surface S_1 will be made to match that of S_2. At the match point, it will

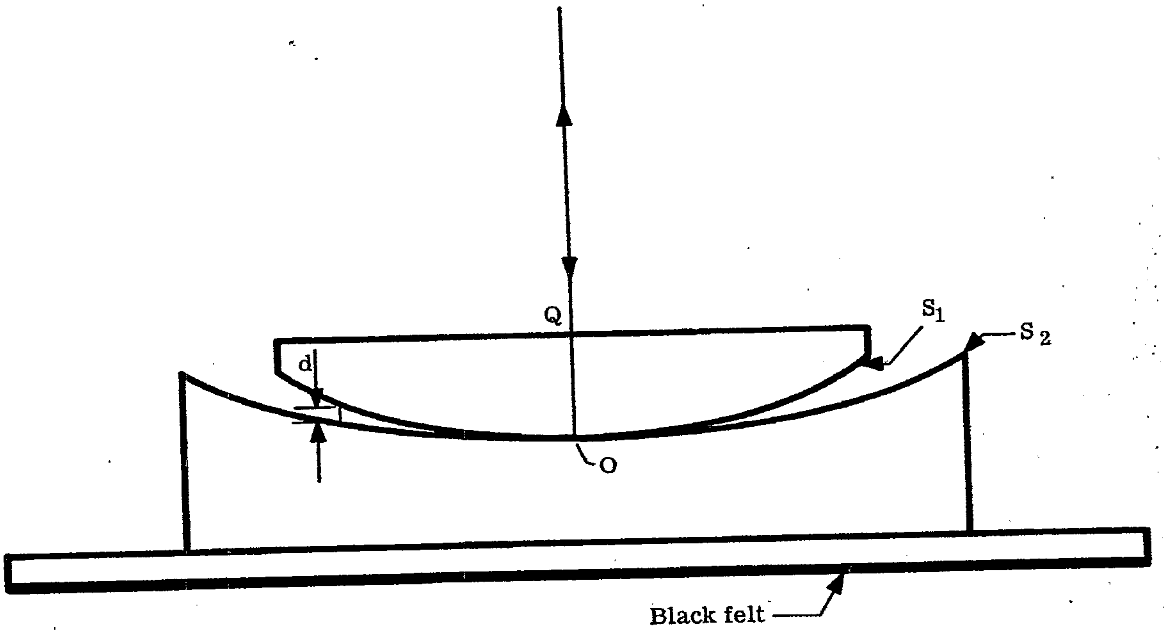


Figure 25.11- Arrangement of the interferometer surfaces S_1 and S_2 for obtaining Newton's fringes.

be possible, as in paragraph 25.8.1.4, to place surfaces S_1 and S_2 into sufficiently close contact so that a single fringe spreads over surface S_1. For work of highest quality, it is customary to specify that this single fringe shall be made uniform in color.

25.8.1.7 The method of the sagitta (see paragraphs 16.13.1.1 and 16.13.1.6) enables one to make a good estimate of the radius of surface S_1 when this radius departs only slightly from that of the test glass. Consequently, it is not always necessary to provide a test glass whose radius is equal to that of the completed surface S_1.

25.8.1.8 When elliptical fringes appear around point O, Figure 25.11, surface S_1 is not spherical. The minor and major axes of the elliptical fringes may be measured, and the ratio of the minor axis to the major axis computed. This ratio is a measure of the ellipticity of surface S_1 and is often utilized as a specification of the maximum tolerable ellipticity. When departures of many fringes from the test glass can be tolerated, another measure of ellipticity is to count the number of fringes along some convenient length in the direction of the major and minor axes of the elliptical fringes and to utilize the ratio of these fringe counts in specifying the tolerable ellipticity.

25.8.1.9 An extreme amount of irregularity in the shape of the fringes is an indication that the tested surface has been improperly polished or molded. "Orange peel" and other defects of polished surface produce irregularities in the observed pattern of fringes.

25.8.1.10 Contrast in the fringes deteriorates as the reflectances a_1 and a_2 of surfaces S_1 and S_2 become more unlike. The light beams reflected from these two surfaces can interfere to produce systems of fringes having zero intensity in the dark fringes (and hence displaying maximum contrast) only when the amplitudes, a_1 and a_2, of the two, coherent, interfering beams are alike. The distribution of intensity in the fringe system when a_1 and a_2 are unlike can be ascertained from paragraphs 16.1.1.3, 16.1.1.5, 16.8.1.1, 16.8.1.2, and 16.9.1.4. In spite of reduced contrast in the fringes, the interferometers of Figures 25.9 and 25.10 are often applied to testing polished surfaces of metals. With metals and other opaque substances, surface S_2 of Figures 25.9 and 25.10 must be that of the opaque material.

25.8.1.11 The reflectance of the "test glass" can be increased by the deposition of a high reflecting coating or by increasing the refractive index of the test glass. In this way, contrast in the fringe system will be improved in testing high reflecting surfaces. Metallic surfaces and coatings produce phase changes on reflection that differ from zero (as in the reflection from a glass-to-air interface) or that differ from $\lambda/2$ (as in the reflection from an air-to-glass interface). Consequently, Equations (3) and (4) require modification. However, the effect of the modified phase changes on reflection is only to shift the location of the fringes. As a result, the interpretation of sections 25.8.1.3 to 25.8.1.9 remains unchanged when applied to coated or metallic surfaces. For testing surfaces of high quality, any coating applied to the test glass must be extremely uniform in thickness and composition because phase changes on reflection can vary appreciably with thickness and composition of the coating.

25.8.1.12 As the reflectance of surfaces S_1 and S_2 is increased from that of polished glass, the nature of the interference fringes formed by the interferometers of Figures 25.9 and 25.10 alters gradually until, finally, these fringes are classified as multiple beam interference fringes. Inter-reflections between surfaces S_1 and S_2 serve to sharpen the fringes formed by the reflected or the transmitted light beams in the manner discussed in paragraph 16.17. With suitable changes in the technique of observation, these sharp (narrow) fringes can be used to detect and measure surface irregularities as small as 10 Angstroms in height. Polished surfaces are found to be rough terrains whose hills and valleys vary in height and depth from 10 to 120 Angstroms. These small irregularities are not visible in the "double beam" interferometer of Figure 25.10 when surfaces S_1 and S_2 are of polished glass.

25.8.1.13 The interferometer method of Figures 25.9 – 25.11 is essentially a "contact" method. Experience and care are required in order to avoid undue scratching of one surface by the other. Life of the test glass is shortened by wear and scratching. A number of convenient interferometer techniques can be applied to testing flat surfaces without placing two surfaces in contact. However, existing interferometer methods for avoiding contact between two spherical surfaces are either so inconvenient to manipulate or so difficult to interpret that the contact method remains the standard method of the optical shop.

25.8.2 The Fizeau Interferoscope.

25.8.2.1 The Fizeau interferoscope, Figure 25.12, is a double beam interferometer that permits one relatively flat surface, S_1, to be tested against another flat surface, S_2, without placing these two surfaces in contact. The increased "working distance", d, is made possible without undue loss of contrast* in the fringes by restricting the effective size of the light source to a pinhole, H, and by illuminating the pinhole with monochromatic light. Since improved monochromaticity and smaller pinholes entail loss of light, the ultimate working distance, d, is restricted by the required level of illumination. Distances of d greater than 2cm must be considered "large" and should be avoided in designing and planning the interferoscope. Fizeau interferoscopes have been varied in design to meet the needs of various users. The use of beamsplitters as illustrated in Figure 16.2 is to be avoided in order to conserve light. The instrument illustrated in Figure 25.12 is an example of one of the more flexible types of interferoscopes. When this instrument is to be used for testing surface S_1 against surface S_2, the two sets of leveling screws, L_1 and L_2, are adjusted in the order mentioned so that the light beams reflected from S_1 and S_2 are refocused by the collimator as images of the pinhole, H, at the aperture, A. When the pinhole images formed at A are brought almost into unison by further relative adjustments on screws L_1 and L_2, straight fringes will appear on the observer's retina as he looks through the aperture, A, provided that the test surface, S_1, is optically flat. Interpretation of the interference fringes remains the same as with the simpler interferometers of Figures 25.9 and 25.10. Except for the more convenient and elegant manner in which the relative inclinations of surfaces S_1 and S_2 can be adjusted with the aid of the leveling screws, the procedures and methods of sections 25.8.1.3 - 25.8.1.5 apply again.

25.8.2.2 Many optical workers use the Fizeau interferoscope exclusively for ascertaining the degree of parallelism of the surfaces of a plane parallel plate. One surface, S_1, of plate, P_1, is first made optically flat. With interferoscopes of the type illustrated in Figure 25.12, test plate, P_2, is removed from table, T_2. The leveling screws L_1 are adjusted so that the two beams reflected from surfaces S_1 and S_{11} are focused within aperture A as images of pinhole, H. If both surfaces of plate P_1 are optically flat but are not quite parallel, the fringes are parallel to the line of intersection of surfaces S_1 and S_{11}. The fringes curve as surface S_{11} departs from flatness. As S_{11} is made optically flat and brought into parallelism with S_1, a single fringe of uniform intensity spreads over the field of view determined by the area of plate P_1. Equations (3) and (4) must be modified to include the refractive index n of plate P_1. Thus dark fringes occur when

$$nd = \nu \frac{\lambda}{2} \; ; \quad \nu = 0, 1, 2, 3, 4, \text{ etc. } ; \qquad (5)$$

and bright fringes occur when

$$nd = \mu \frac{\lambda}{4} \; ; \quad \mu = 1, 3, 5, 7, \text{ etc.} \qquad (6)$$

---

* See paragraphs 16.4 and 16.5 for the effect of monochromaticity, pinhole size and separation, d, on fringe contrast. Other principles underlying the use and interpretation of the Fizeau interferoscope are discussed in 16.2 and 16.2.2.

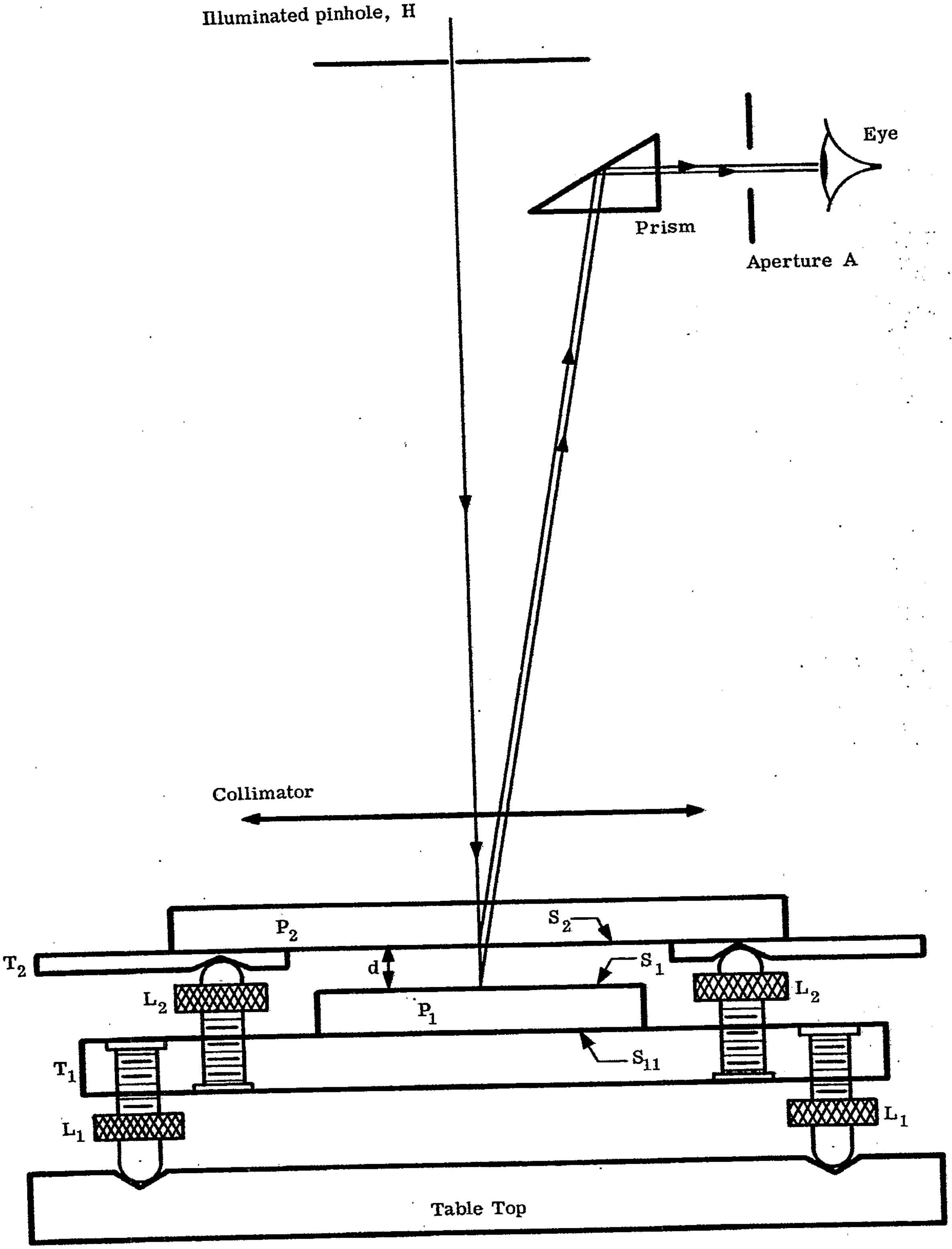


Figure 25.12 - A Fizeau interferoscope.

In other words, dark fringes occur when the optical path nd of the plate is equal to any integral number of half wavelengths, and bright fringes occur when the optical path is an odd number of quarter wavelengths. These conclusions can be expected intuitively when one considers that the phase change on reflection will be $\lambda/2$ at surface S_1, and 0 at surface S_{11}, so that the difference in the phase changes on reflection is $\lambda/2$. The beam reflected from S_{11} passes through plate P_1 twice, and thus is increased in phase by twice the optical path or 2(nd). Because the phase change on reflection is $\lambda/2$ at the air-to-glass interface, S_1, the two interfering beams proceed toward the observer with a phase difference, Δ, given by

$$\Delta = 2nd - \lambda/2. \qquad (7)$$

If now, nd is given by Equation (5), $\Delta = (\nu - 1/2)\,\lambda$ so that destructive interference takes place. But if nd is given by Equation (6), $\Delta = (\mu - 1)\,\lambda/2$. Since $\mu - 1$ must be an even number, $(\mu - 1)/2$ is an integer and Δ is an integral number of wavelengths. Hence we verify that constructive interference takes place when the optical path obeys Equation (6).

25.8.2.3 As an example of the sensitivity of the Fizeau interferoscope in testing for parallelism of the two surfaces of a plate, let us suppose that the diameter of the plate is 5cm, that its refractive index is 1.5, that the wavelength λ is 0.5461×10^{-3} mm and that the optically flat surfaces define a wedge whose optical path differs by $\lambda/2$ at the extreme ends of the wedge. Suppose that a bright fringe appears at the thin edge of the wedge. A bright fringe must appear at the thick end of the wedge since the optical path is greater by $\lambda/2$ at the thicker end of the wedge. This conclusion follows at once from Equation (6); for if nd is increased by $\lambda/2$, μ is increased to the next odd number, $\mu + 2$, the spectral order number of the next bright fringe, Equation (5) will be satisfied at the center of the plate so that a dark fringe occurs here. The field of the plate will appear very nonuniform. It presents one dark and two bright fringes. Despite this nonuniformity, the angle, α, between the surfaces of the plate is less than one second of arc. Since nd changes by $\lambda/2$ across the plate, the thickness of the plate changes by $\lambda/2n$. Therefore

$\alpha = \dfrac{\lambda/2n}{\text{diameter}} = \dfrac{0.5461 \times 10^{-3}}{3 \times 50} = 3.64 \times 10^{-6}$ radians or 0.75 seconds of arc. If the variation of intensity across the plate is reduced to 0.1 fringe, α will be reduced to 0.075 seconds.

25.8.3 A Modified Michelson Interferometer.

25.8.3.1 A flexible instrument, with the aid of which any surface (whether glass or metallic) can be tested for flatness against an optical flat without contact, is illustrated in Figure 25.13 as a specialized form of Michelson's interferometer. If S_1 is a surface of polished glass, surface S_2 is chosen as polished glass. If surface S_1 is metallic or high reflecting, an optical flat P_2 having surface reflectance approximating that of S_1 will be provided. The user can afford to supply several optical flats since these flats will not have to be replaced because of wear and scratches. The housing, H, is compact and rigid. It is designed to support the beamsplitter and the optical flat, P_2, with minimum vibration. The line OB is pointed in the vertical direction so that the test plate, P_1, is simply laid upon an auxiliary, stable support, Q. The arms OB and OA of the interferometer will be designed so that these arms are nominally of equal length and so that these arms are easily adjusted for equal lengths. Adjusting screws L can be utilized both for equalizing the lengths of the arms and for tilting surface S_2 with respect to S_1 to control the fringe width. The mechanism for tilting plate P_2 must be designed with great care because it is this mechanism that determines the operator's convenience in making quick and certain adjustments of the fringe widths as well as in checking arms OB and OA occasionally for equality by finding the "white light position". Supports Q should be ground to equal thicknesses in order to avoid hunting for the white light position each time Q is replaced by another support. An auxiliary, tiltable mirror M is provided for deflecting the light beam toward the observer.

25.8.3.2 Michelson's interferometer does not differ in principle from the interferometers of Figures 25.9 and 25.10. Consequently, the conclusions of paragraph 25.8.1.2 remain valid and the methods of paragraphs 25.8.1.3 - 25.8.1.5 apply again. Figure 25.14 illustrates why the Michelson interferometer behaves as the interferometers of Figures 25.9 or 25.10.

25.8.4 4 The Twyman-Green Interferometer.

25.8.4.1 The Twyman-Green interferometer is similar to the Fizeau interferoscope as regards basic principles* and interpretation. In both of these interferometers, the allowable optical path difference between the two interfering waves is increased by reducing the effective size of the source to that of a pinhole and by increasing the monochromaticity of the source of light. If, for example, mercury arcs are utilized, they should be operated at reduced pressure and followed by a high quality filter for 5461 Angstroms. The Twyman-Green interferometer has many points of mechanical similarity with Michelson's interferometer. However, Michelson's interferometer is invariably intended for use with broad sources of light.

---

* The principles underlying the Twyman-Green interferometer are discussed in paragraph 16.3.

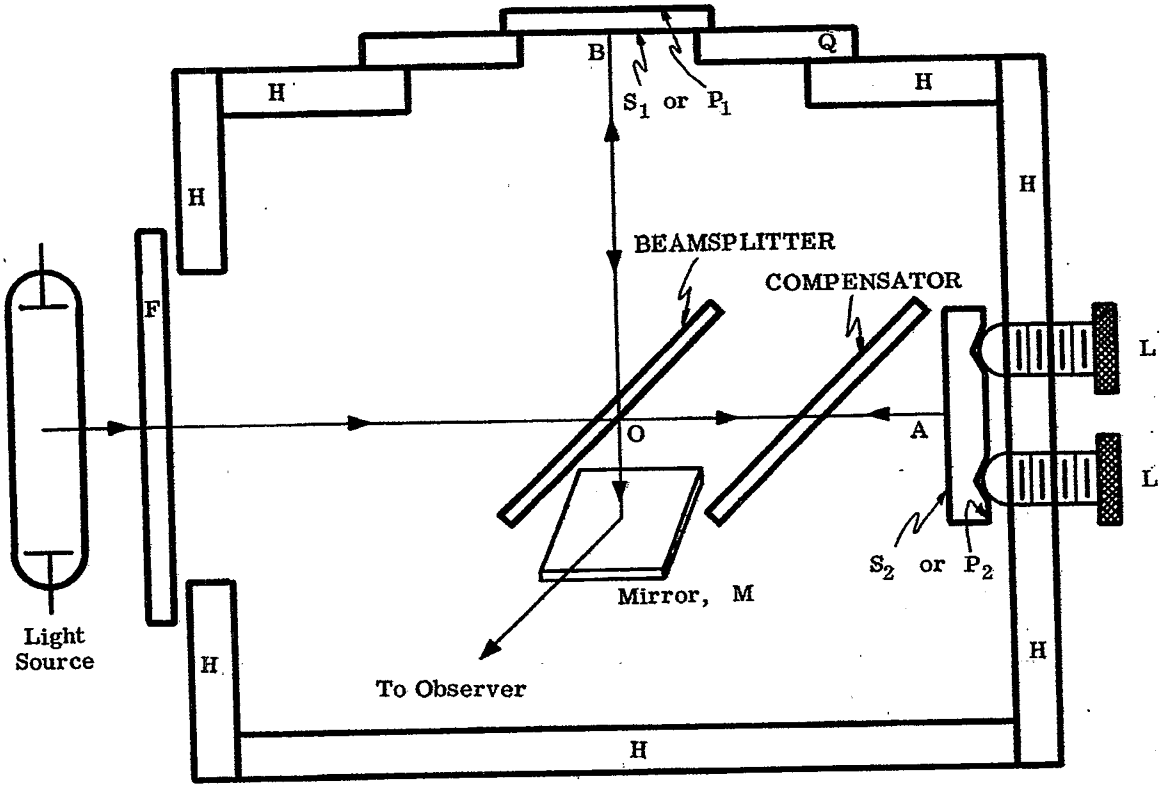


Figure 25.13- A modified Michelson interferometer.

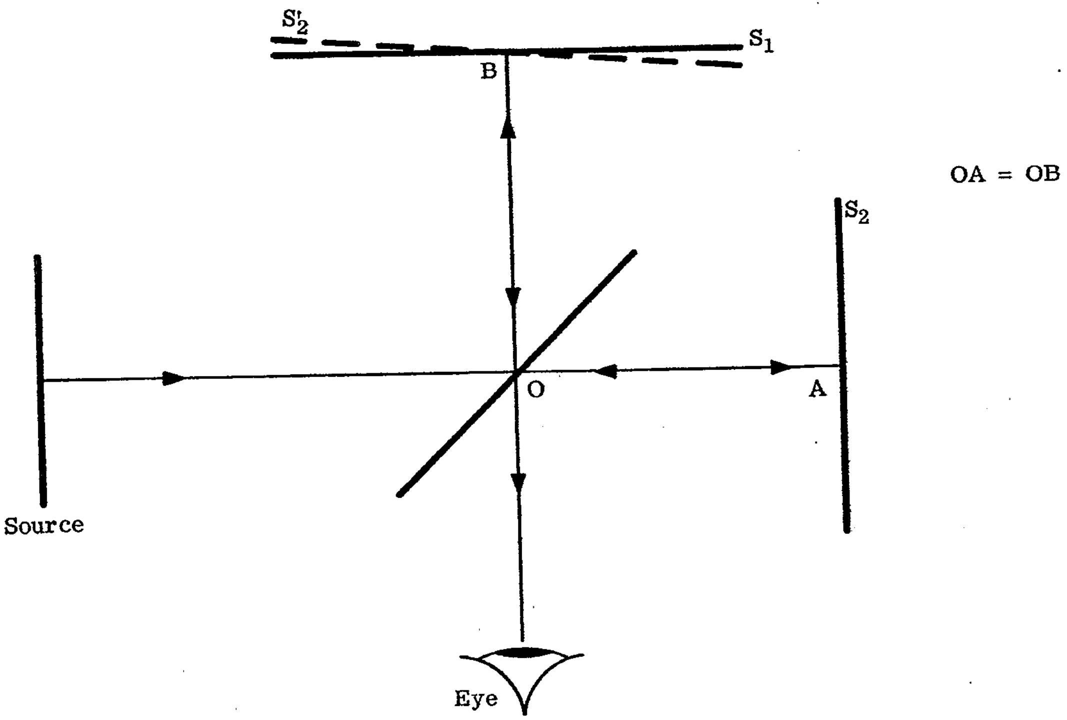


Figure 25.14- Michelson's interferometer as a method of Fizeau fringes. The observer sees surface S_2 as though it were located at S'_2, consequently, the fringes are formed as by reflection from two surfaces, S_1 and S'_2, that are in close contact.

25.8.4.2 Emphasis in the design of the Twyman-Green interferometer is placed upon taking advantage of the permissibly large optical path difference, d, between the two arms of the interferometer in measuring the variations of optical paths through plates or prisms. The instrument is not intended for regular use in checking flatness of surfaces. For maintaining best contrast in the fringes, the optical paths OA and OB, Figure 25.15, should be kept approximately equal. Accordingly, end mirror, M_2, is mounted on a slide that permits M_2 to be moved without appreciable wabble along the line AO. An iris diaphragm whose opening can be reduced to 0.75 mm, or less, in diameter is ordinarily used as pinhole, H. By means of this adjustable iris, the observer can choose his own compromise between brightness and contrast of the fringes. End mirrors, M_1, and M_2, are adjusted by means of suitable mechanisms (not shown) involving screws, L_1 and L_2, until pinhole images of H, formed after reflection at the end-mirrors, appear within aperture, A, at the rear focal plane of the telescope. If these pinhole images are not too far apart within A, fringes will be seen when the eye is placed behind A. One obtains the desired fringe width by further adjustments of L_1 and L_2. Failure to obtain good fringes as the pair of pinhole images is brought into unison indicates that the optical path difference between arms OA and OB is too great.

25.8.4.3 Figure 25.15 illustrates the arrangement for testing plates, P. Suppose that one knows that the surfaces of the plate are optically flat and that he wishes to check the uniformity of the optical paths through the plate. These optical paths can differ due to nonparallelism of the surfaces or due to nonuniformity in the refractive index, n. In one procedure the end mirrors, M_1 and M_2, are adjusted so that one fringe spreads over the field of view before plate, P, is inserted. The effects of introducing plate, P, are then observed. If the broad fringe is left practically undisturbed, the optical paths through the plate are sensibly uniform. The appearance of straight fringes indicates that the surfaces of the plate are not parallel. Irregularities in the fringes indicate that the refractive index is not constant. Let Δt and Δn denote variations in the thickness, t, and refractive index, n, of the plate, respectively, and let Δd denote the corresponding variation in the optical path difference, d, between the two arms of the interferometer. If Δn is negligible,

$$\Delta d = 2(n - 1) \frac{\Delta t}{\lambda} \text{ wavelength numbers.} \tag{8}$$

If Δt is negligible,

$$\Delta d = 2 \Delta n \frac{t}{\lambda} \text{ wavelength numbers.} \tag{8a}$$

The factor, 2, enters because light waves traverse the plate twice. The following principles should be kept in mind.

(a) Each fringe is associated with a particular value of Δd.

(b) When one of the arms of the interferometer is altered in length by pressing upon the plate that supports the elements of the interferometer, each fringe moves such that Δd remains constant. With respect to the problem of interpreting the fringe system for the direction of the wedge that produces the straight fringes, one has only to shorten or to lengthen one arm of the interferometer and to note the corresponding movement of the fringes.

(c) In passing from one fringe to the next fringe of equal darkness or brightness, Δd changes by unity.

(d) In passing from a bright fringe to the next dark fringe, Δd changes by 1/2.

25.8.4.4 The base plate of the Twyman-Green interferometer is grooved or otherwise constructed so as to permit the end mirror, M_1, to be swung into orientations for testing prisms, etc. The configuration for testing right angled prisms is illustrated in Figure 25.16. Mirror, M_1, is adjusted for the desired fringe width. If all surfaces of the prism are known to be optically flat, the observed fringes reveal the degree of uniformity of the refractive index of the prism. More frequently, one will not have explored independently the degree of flatness of the surfaces of the prism. In such cases the observed fringes reveal the combined effects due to departures from surface flatness and due to inhomogeneities in refractive index. This method does not appear to have been modified to yield information about the angles of the prism or about deviation by the prism.

25.8.4.5 The following expedient is used, as the occasion demands, for distinguishing between effects due to inhomogeneity of refractive index and due to inadequate flatness of surface. The method is particularly effective in testing plates. As illustrated in Figure 25.17, plate, P, of Figure 25.15 is "immersed" between two optically homogeneous plates, E and F, that have outer surfaces of high degree of flatness. These plates are preferably, but not necessarily, plane parallel plates. For best results, the refractive indices of the immersion oil and of plates E and F should match the refractive index of the test plate, P. It can be useful to

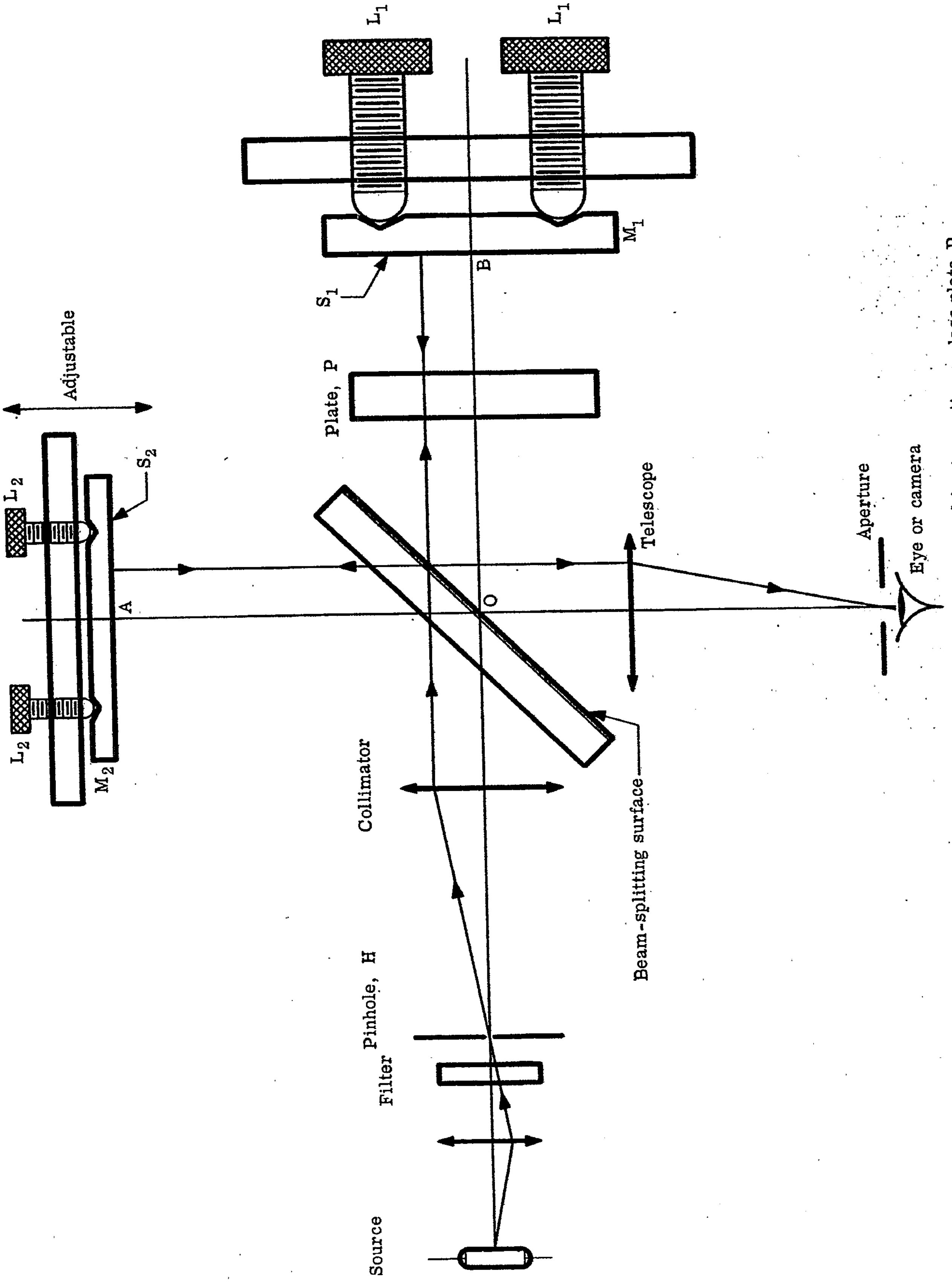


Figure 25.15- Twyman-Green interferometer as used for inspecting a glass plate P.

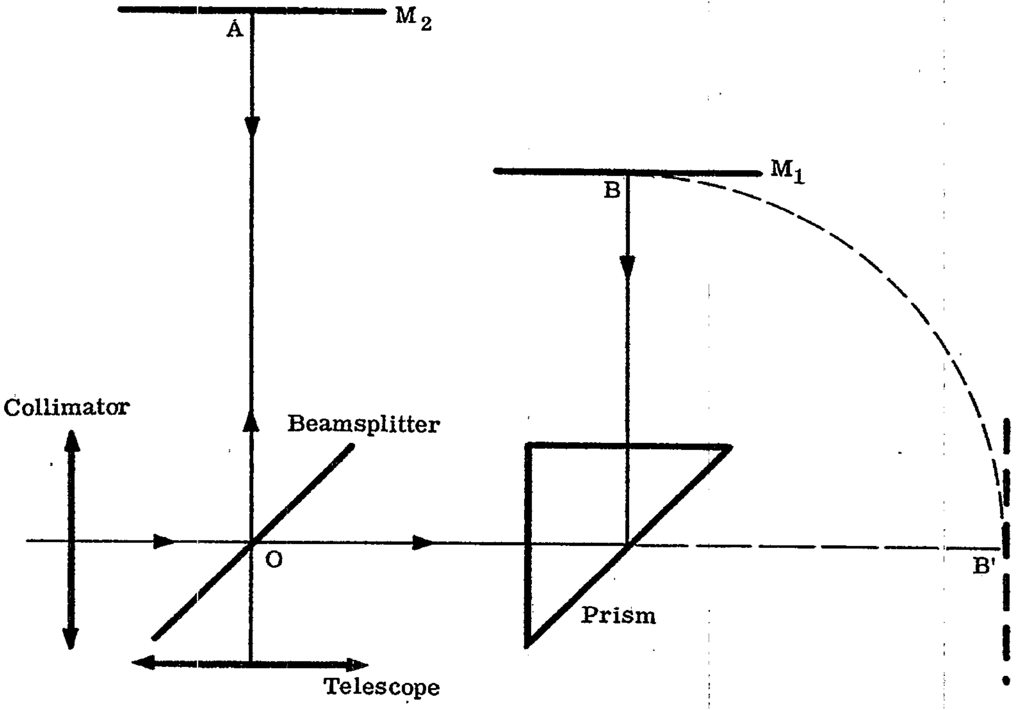


Figure 25.16 - Modification of the location of end mirror M_1 for testing prisms in the Twyman-Green interferometer.

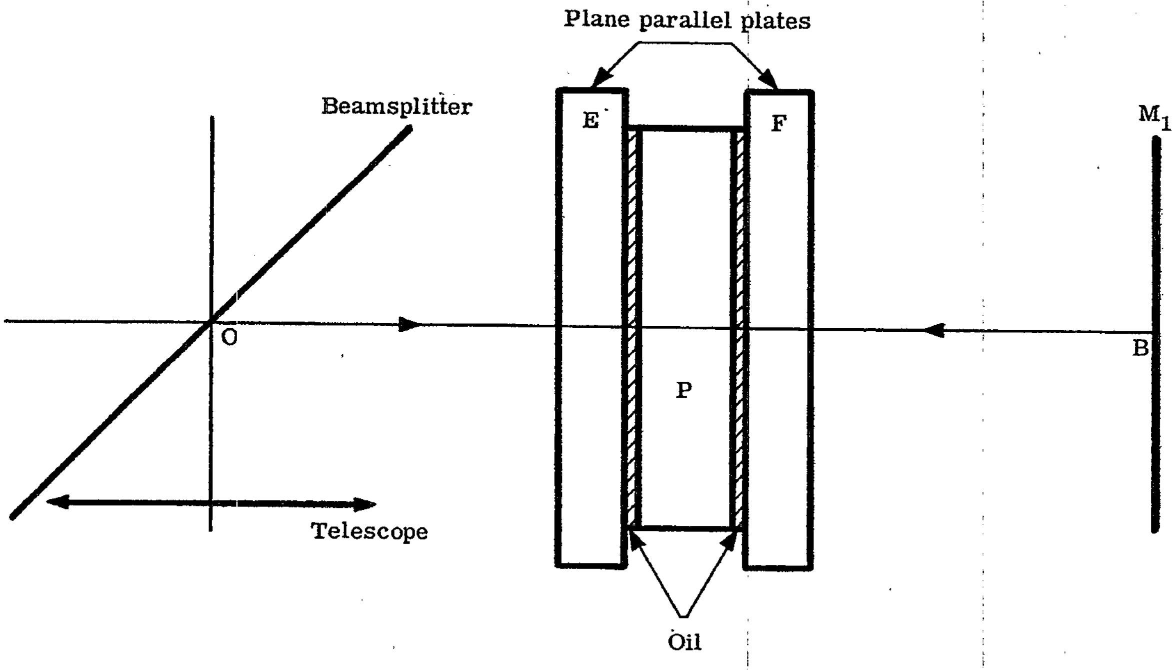


Figure 25.17 - Method for obviating surface flatness of the test plate P.

"immerse" the prism of Figure 25.16 between two plane parallel optical flats. However, it is not possible to avoid lack of flatness at the reflecting surface of the prism.

25.8.4.6 Twyman-Green interferometers are provided, as illustrated in Figure 25.18, with an accessory fixture for observing spherical aberration of objectives throughout a wide range in focal lengths. Suppose that neither the collimator nor the test objective L possesses spherical aberration. Then rays from the axial point within pinhole, H, will be converged upon the axial point, C, at the rear focal plane of lens, L. If a spherical, convex mirror, M_1, is placed as indicated with its center of curvature at point C, all rays, ef, will be reflected back upon themselves and will emerge from lens, L, as normals to a plane wave that is propagated toward the beamsplitter. This plane wave interferes with the plane wave reflected from M_2 in the other arm of the interferometer to produce upon the observer's retina a family of straight fringes whose fringe width depends upon the angular adjustment of the end-mirror, M_2. In particular, a single fringe can be spread over the field of view. If the test objective has spherical aberration, all rays, ef, cannot be reflected back upon themselves. Consequently, the wave emerging from objective, L, will not be plane and will interfere with the plane wave from the other arm of the interferometer to produce interference fringes that display axial symmetry provided that the elements of objective, L, have axial symmetry and are well centered.

25.8.4.7 With objectives having long focal lengths, it is preferable to locate the convex mirror M_1 near the objective. This means that spherical reflectors having a series of radii should be provided. With objectives having short focal lengths, such as microscope objectives, the available working distance will not permit a convex reflector, M_1, but rather a concave reflector, M_1', centered about point C must now be used.

25.8.4.8 In actual practice, the exact location of the center of reflector, M_1, with respect to the axial point near the rear focal plane of lens, L, is problematical and becomes often a matter of choice. The spherical aberration can be less with respect to a focal plane that falls on one side or the other of the paraxial focal plane. Secondly, some consideration will show that the observed spherical aberration is not necessarily that of the objective alone. When the test objective has spherical aberration, the observed aberration is, in fact, the aberration of the combination consisting of the test objective, L, and the spherical mirror, M_1. Interpretation of the fringe displacements for the spherical aberration of the objective alone is not without objections of fundamental nature. But inspite of this difficulty, the Twyman-Green interferometer method is one of the better methods for indicating actual or relative amounts of spherical aberration in objectives of high optical quality.

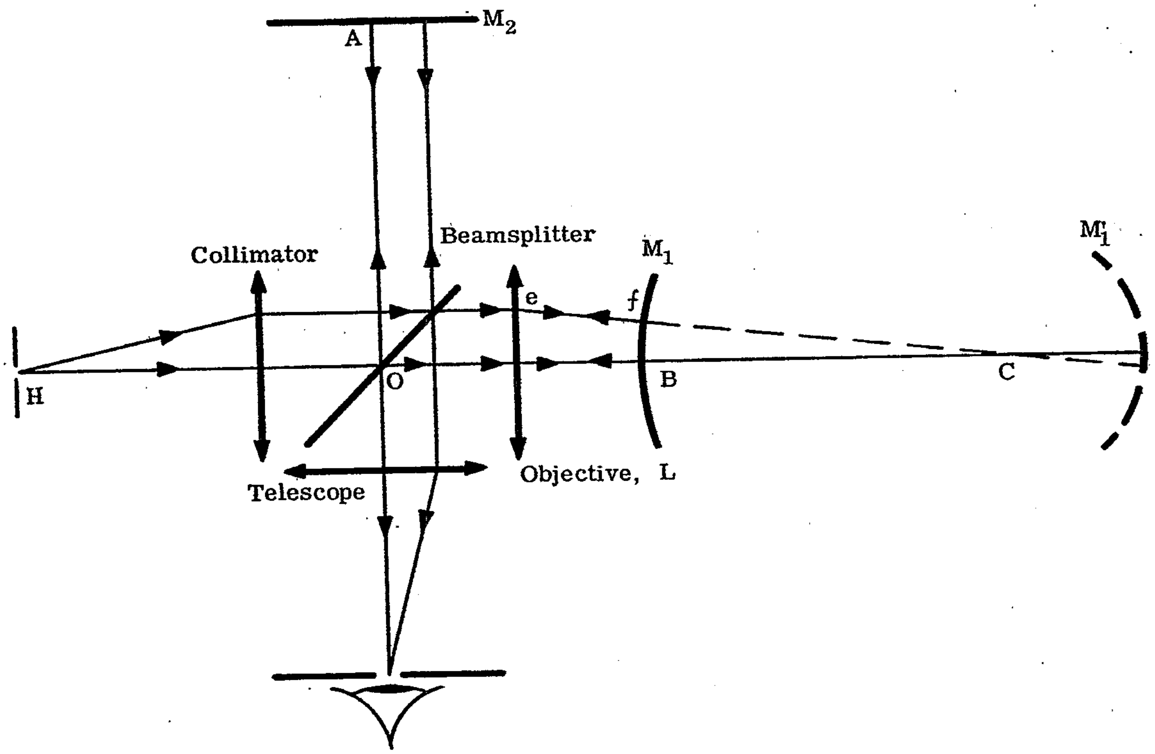


Figure 25.18- Adaptation of the Twyman-Green interferometer to the observation of spherical aberration of lenses, L.

25.9 THE RONCHI TEST

25.9.1 Introduction. One of the fairly common tests used currently by optical workers is the Ronchi Test(17). This technique actually falls into a class known as the "shadow-fringe method". The geometrical aspects of the technique are outlined in Figure 25.19.

25.9.2 Theory.

25.9.2.1 The essence of the theory is as follows. Let us suppose as a start that a lens, L, produces a perfect star image at the paraxial focus, O, Figure 25.19. (The edge ray shown will not pass through B but through O.) Now if a plane grating of 4-8 lines/mm (Jentsch method 18 and 19) or 10-20 lines/mm (Ronchi method) is positioned at an arbitrary distance, g, from the focal point, O, as indicated in Figure 25.19, a shadow of it would be formed on a screen placed in a plane of projection arbitrarily distant p from O. Since we are, for the present, considering a perfect lens, rays from all parts of the lens aperture pass through O and this point is the single center of projection of the grating onto the plane of projection. Hence the shadow image of the grating has exactly the proportions, over all its area, of the grating itself, i.e. without distortion and with a constant magnification over its area of

$$\frac{\sqrt{\eta + \xi}}{\sqrt{y + x}} = \frac{p}{g} \qquad (9)$$

This ideal aberrationless case is usually not found and there is some spherical aberration, S, (shown for the edge ray in the figure) with the intersection points of other rays filling the space between O and B (simple undercorrection). Thus there is no longer a single point which can be considered as the center of projection for the entire grating. In this case the shadow image of the grating cast on the plane of projection will have a varying size scale over its area because centers of projection for points on the entire area of the grating lie between B and O. The magnification for any point on the grating becomes a function of the spherical aberration, S, of the ray passing through that point in accordance with the expression

$$\frac{\sqrt{\eta + \xi}}{\sqrt{y + x}} = \frac{p - S}{g - S} \qquad (10)$$

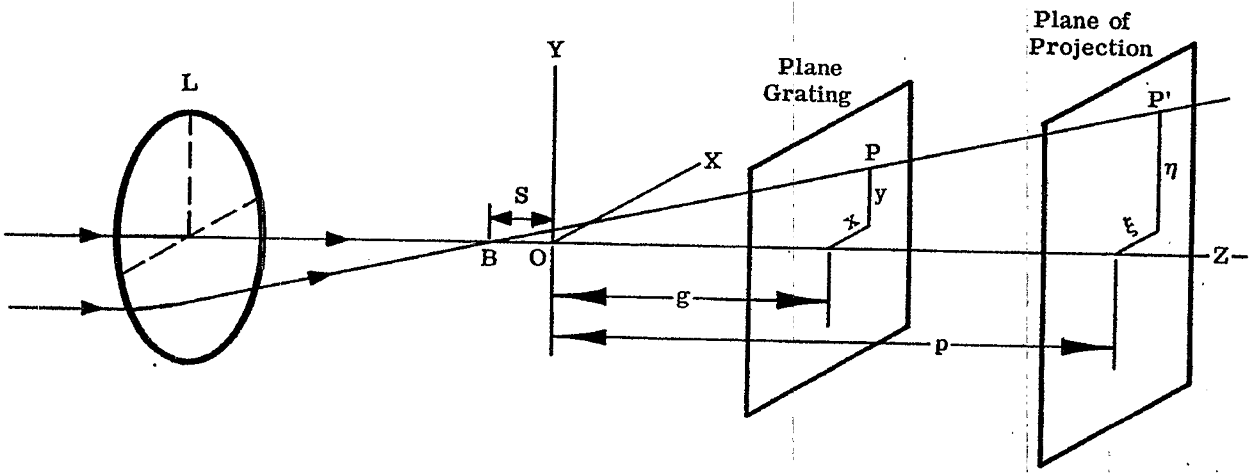


Figure 25.19-Geometrical theory of the Ronchi Test.
(After Martin's, Technical Optics, Vol. II, Pitman Pub. Co. 1950)

---

(17) Ronchi, Ann. d. R. Scula Normale Supriore di Pisa, Vol. XV (1923)
(18) Jentsch, Physikal Zeitschr, XXIX, 66, (1928)
(19) Martin, Technical Optics, Vol. II, 289, Pitman, 1950

If the grating is composed of straight lines its shadow image will show the lines as curved, and from the geometry of the figure, it can be shown that the curvature indicates the amount of spherical aberration present.

25.9.2.2 The complete theory must take into account not only the geometrical aspects outlined briefly here, but also the fact that interference may well be significant if the grating is as fine as those used by Ronchi or if more definitive interpretations are required. F. Toraldo Di Francia (20) has treated this subject in quite some detail, and the interested reader is urged to consult his paper.

25.9.3 Ann Arbor Tester.

25.9.3.1 There has appeared a commercial unit known as the Ann Arbor Tester based on this principal and manufactured by the Ann Arbor Optical Co. The tester is the device on the right of Figure 25.20.

25.9.3.2 The following Figures 25.21 and 25.22 are taken from the instruction booklet furnished with the instrument. Figure 25.21 [4(a)-(g)] shows the testing of an eight-inch focal length spherical mirror as a 175 line/inch grating is moved through focus on axis. Figure 25.21 [(5(a)-(d)] shows the patterns obtained with the Optical Tester and a thirty-inch focal length paraboloid cut 7° off axis. The pattern in [5(a)] shows the tester on the optical axis while [5(b)] shows the tester off axis; [5(c)] and [5(d)] were taken in the same positions but the parabolic mirror was rotated. The experimental arrangement is shown in Figure 25.22.

25.9.4 Jentsch's grid method. A similar testing technique using the coarser gratings of Jentsch is shown in Figure 25.23 taken from Martin (21) and showing the presence of spherical aberration. One optical shop checks all its work with this technique finding it a very sensitive and simple method for examing mirrors and lenses. The shadow - fringe method has much in its favor as an experienced worker gains a feeling quickly as to the nature of the defects of the system under test.

25.9.5 Summary. In the last analysis, however, all optical tests depend upon evaluation, and the experience of the optical worker himself is a vital factor. It is largely a matter of what the workers in a particular laboratory have previously used. One of the largest laboratories in the United States does practically no Ronchi-type testing as they have accumulated other equipment and know-how over the years that gives them the information they need.

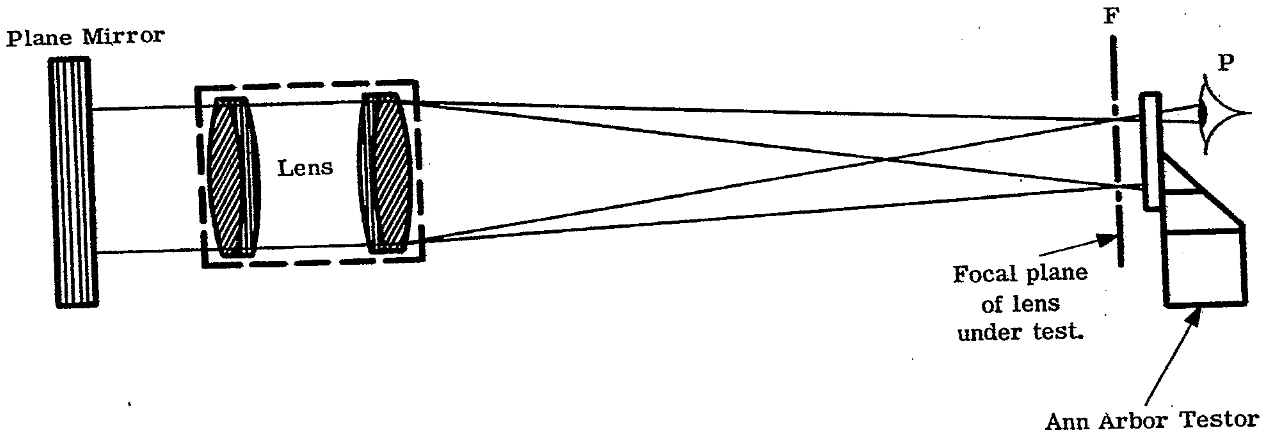
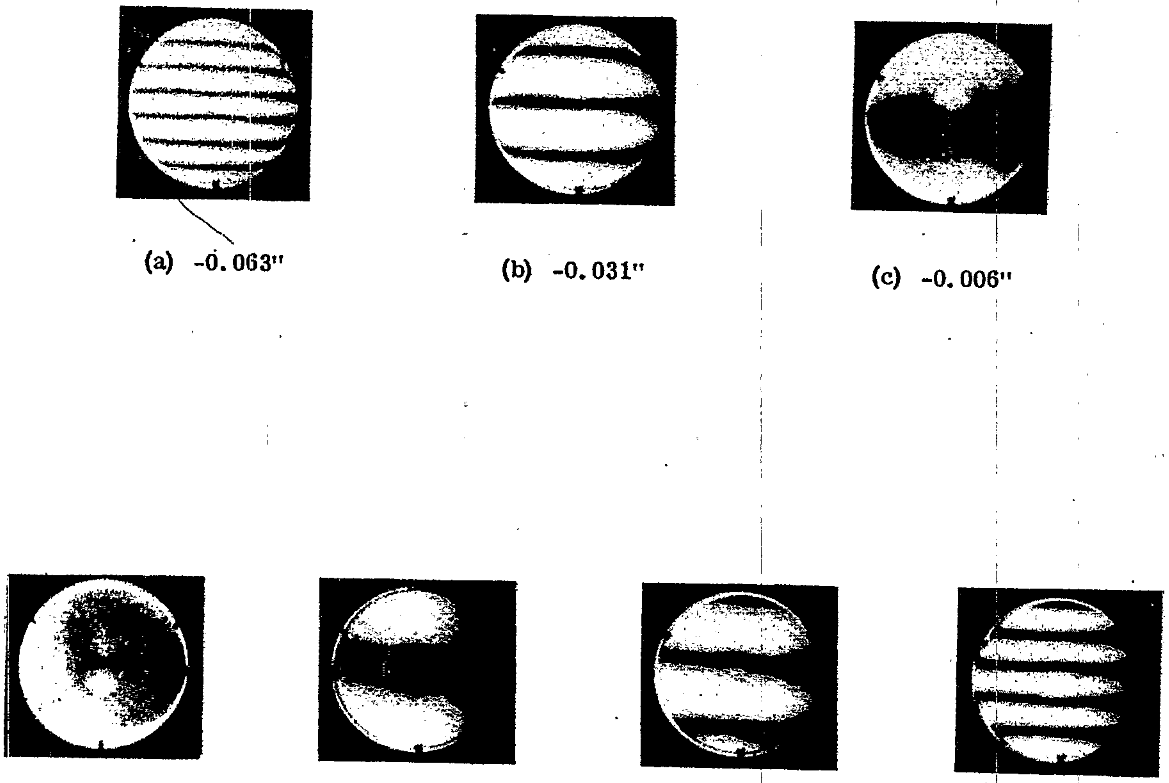


Figure 25.20- The Ann Arbor Tester.

---

(20) di Francia, Optical Image Evaluation, (NBS Circular 526), 161, U. S. Gov't. Printing Office, (1954).
(21) Martin, Technical Optics, vol. 1, p289, Pitman, (1948). also Jacobs, Fundamentals of Optical Engineering, McGraw Hill, (1943).



(a) -0.063"

(b) -0.031"

(c) -0.006"

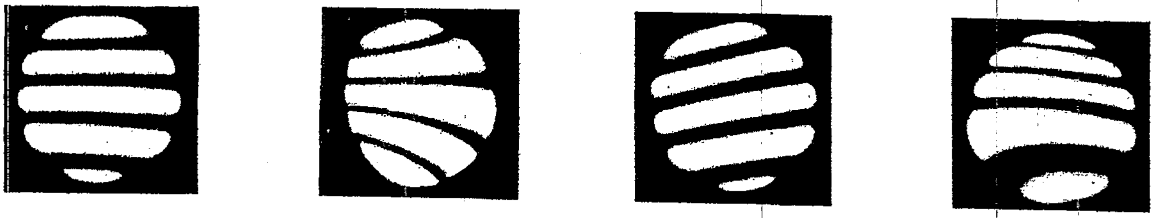
(d) 0.000"

(e) 0.008"

(f) 0.025"

(g) 0.051"

Figure 4



(a)

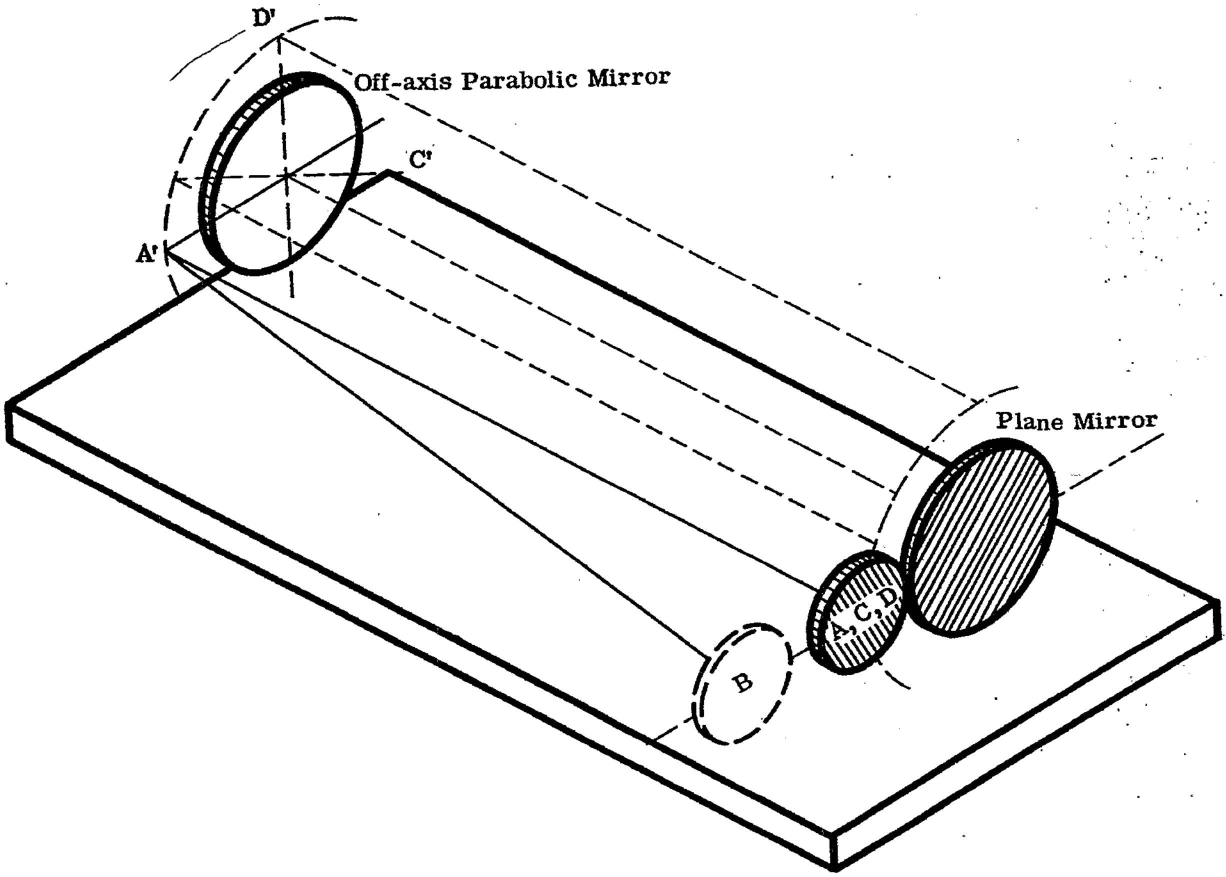
(b)

(c)

(d)

Figure 5

Figure 25.21- Patterns seen with the Ann Arbor tester for various optical systems.



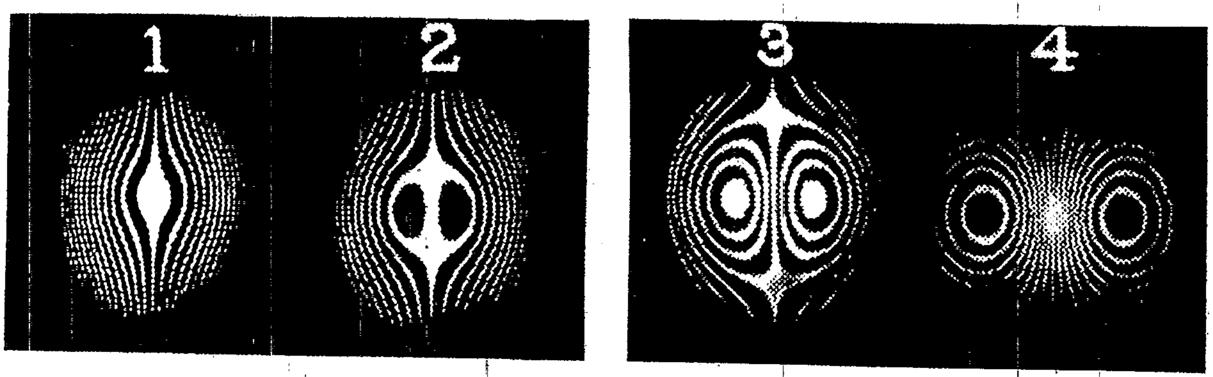
DETERMINING THE OPTICAL AXIS OF OFF-AXIS PARABOLIC MIRRORS

Often off-axis parabolic mirrors are received from the fabricator without markings indicating the side of the mirror which is toward the optical axis. Also, even with such markings, only the plane containing the axis is defined, and it is still necessary to locate the axis.

The position of the optical axis is easily determined by observing the pattern obtained with the Optical Tester. Only when the Tester is on the optical axis will the fringes be equally spaced and parallel to each other and to the lines in the grating when using an arrangement as shown in Figure 25.20.

The resulting patterns for the Tester in the correct and other positions with respect to the axis can be seen from the patterns pictured in Figure 25.21. Figure 5(a) in 25.21 shows the pattern obtained with the Tester on the optical axis of a 30" focal length parabolic mirror cut 7° off-axis. This position is illustrated above, with the Tester at A and the edge of the parabolic mirror closest to the optical axis at A'. Figure 5(b) in 25.21 shows the pattern when the Tester, parabolic mirror, and plane mirror are still in the correct plane, but the grating is located outside the optical axis (B in the Figure above).

Figure 25.22- Determining the optical axis of off-axis parabolic mirrors with the Ann Arbor Tester.



1 and 2 are shadow fringes outside the caustic.

3 and 4 are shadow fringes inside the caustic.

The appearances follow in the progression 1, 2, 3, 4.

Figure 25.23-Jentsch's grid method.

(From Martin's, Tech. Optics, Vol. II, Pitman Pub. Co. 1950)

25.10 FOUCAULT TEST

25.10.1 Introduction.

25.10.1.1 Having been given a lens or mirror surface prescription and having ground and polished the surfaces by hand or machine methods, the question arises as to what areas need to be "figured", i.e. be repolished to achieve perfectly the prescription. While helpful, the viewing of the image in toto is of less value than might be supposed, in that it represents the summation of the contributions from all parts of the surface. A technique which allows for inspection of the surfaces themselves is obviously required. One of the simplest and yet most delicate of all such techniques of surface testing was developed by Foucault (22) in 1859. The method requires, in its simplest form, merely a pinhole, a knife edge, the lens or mirror, and the eye of the observer. The system is shown diagrammatically in Figure 25.24 for a mirror.

25.10.1.2 In essence, the pinhole provides a small source of light which illuminates the surface of the mirror but which is so shielded that it sends no light directly into the eye. If we assume that the spherical surface is perfect and that the longitudinal aberration is negligible, then all rays striking the mirror will be focussed at some point, F. If the pinhole is located at the center of the curvature of the system, then F also will be at the center of curvature.

25.10.1.3 Assuming that the eye is placed close enough to the image so as to view the mirror in the Maxwellian sense (i.e. the eye receives all the rays coming to the focus), then the mirror surface will be evenly illuminated. (Actually even a perfect surface will show some deviations which will be discussed later.) If now a knife edge K is advanced in the direction indicated, the mirror surface will appear to go from completely bright to completely dark as the knife passes through F. Again for reasons to be discussed later this does not quite happen. If the knife edge is displaced toward the mirror from F then as it cuts into the beam the lower side of the mirror is darkened gradually and not until the beam is completely occulted will the eye see no light. If the knife edge is displaced from F away from the mirror, the reverse occurs. Clearly then the point where the intensity varies most rapidly with knife edge movement from bright to dark is the focus. If the knife edge always remains in the plane of the pinhole, then there will be only one place where the mirror darkens uniformly and rapidly with lateral displacement of the knife edge. This point is of course the center of curvature.

25.10.1.4 The extreme delicacy of this measurement will become more obvious if we study Figure 11.28 where a highly exaggerated error is present. Suppose there is an error of slope, angle, θ. The rays hitting

---

(22) Foucault, Ann de L'Obs. de Paris, V, 197 (1859).

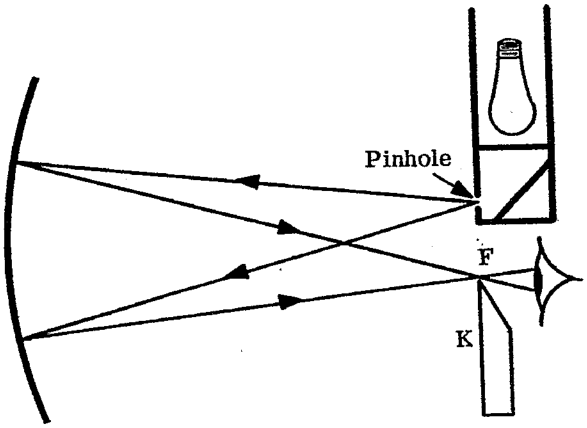


Figure 25. 24 - Experimental arrangement for a Foucault Test of a mirror.

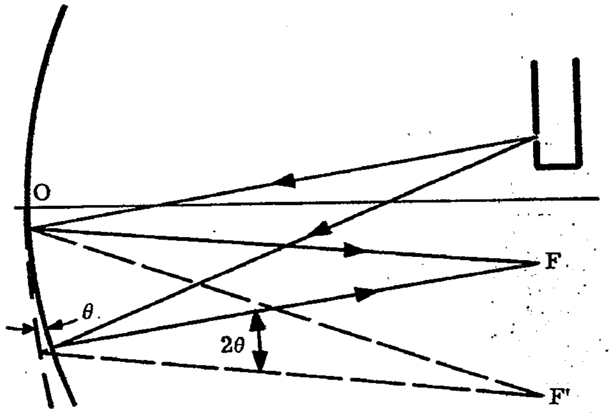


Figure 25. 25 - Schematic demonstration of sensitivity of Foucault testing.

this area will then be deflected from the correct focus by an angle of 2θ and the focal point for this area will be moved laterally a distance approximately equal to FO x 2θ. If the surface has a radius of curvature of ten and a slope angle error of 10^{-5} radian, then the deflection is 10^{-4} inches or .025 mm. An error this size on a lens of such small curvature would be barely detectable but on a larger focal length system it would be clearly visible. The test is actually so sensitive that slope angle errors of 10^{-6} radians are easily seen on the long focal length mirrors used in some telescopes.

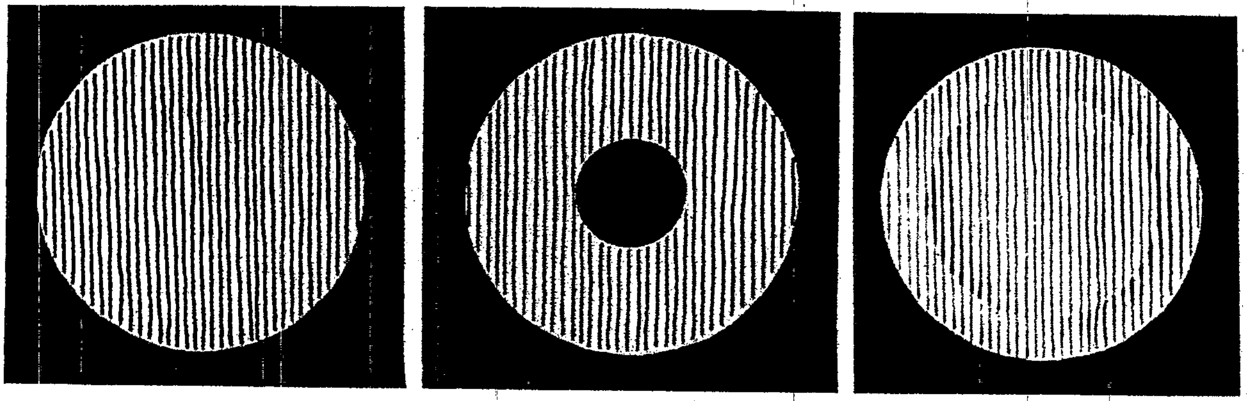
25.10.1.5 Errors of this type usually appear as zones on the surface rather than isolated areas. Unfortunately while Foucault tests are very common, they are almost always done visually and few photographs are taken. The photographs of some drawings from Strong (23) are shown in Figure 25.26. The artistry of Roger Hayward, illustrator for Strong, clearly shows the variations in mirror illumination produced under the Foucault Test.

25.10.2 Detailed discussion of Foucault Test for spherical surfaces.

25.10.2.1 The preceding discussion has been highly qualitative and obviously over-simplified in some respects. To begin with the focus is never a pure geometrical point as we have demonstrated earlier. Secondly, the surface of even a "perfect" mirror does not appear all bright or all dark. It has been known for a long time that at the edge of a circular mirror there appears a very bright ring even when the knife edge has apparently cut through all of the rays. Banerji (24) has also observed that the surface for a real system with finite focal area does not grow continuously darker as the knife edge advances but rather the entire surface presents large variations in illumination. The peak illuminations get smaller and smaller until finally the whole surface is dark. Lord Rayleigh (25) attempted to explain the first of these two effects and was relatively successful. It remained for Zernike (26), Gascoigue (27), and recently Linfoot, in his articles and more recently in his book (28), to carry the interpretation of the Foucault patterns into the realm of the quantitative. Linfoot shows that the patterns may be determined analytically for an aberration-free system by assuming that electromagnetic waves originate at the surface being tested and combine according to the usual interference principles. In the event that the system is not aberration free, one must assume that the electromagnetic waves start at the pinhole and are reflected from the surface in the usual way.

---

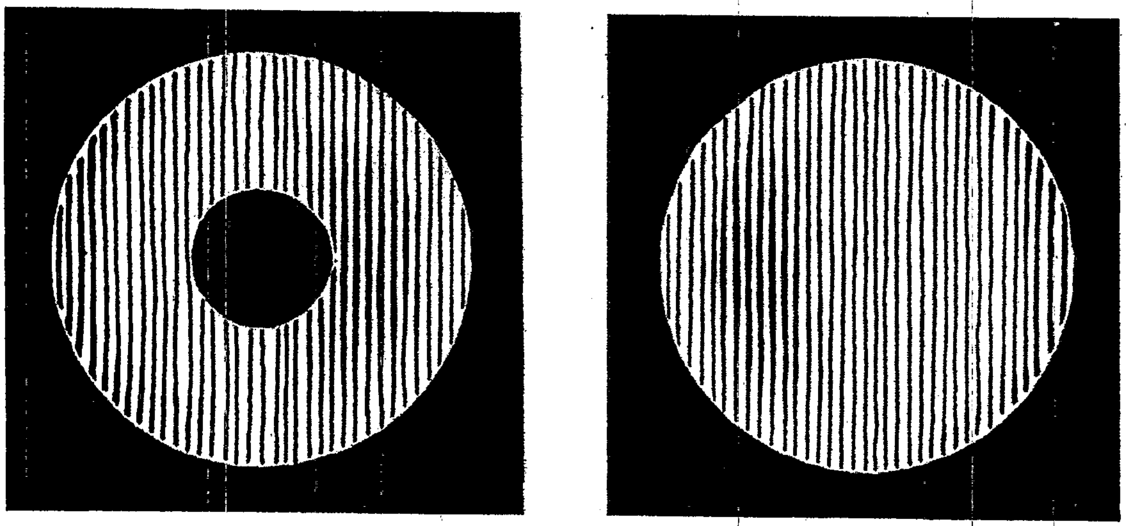
(23) loc. cit., (7), 296, 297
(24) Banerji, Astrophysical Journal 48, 50, (1918)
(25) Rayleigh, Phil. Mag. 33, 161, (1917)
(26) Zernike, Physica 1, 689 (1934)
(27) Gascoigue, M. N. 104, 326, (1945)
(28) Linfoot, Recent Advances in Optics, 128 at req. (1955) Oxford



(a) Spherical mirror tested at the center of curvature.

(b) Parabolic mirror tested with a flat testing mirror.

(c) Spherical mirror with a raised annular ridge as tested at the center of curvature.



(d) Spherical mirror tested with a flat testing mirror.

(e) Parabolic mirror tested at the mean center of curvature.

Figure 25.26-Foucault test appearances.
(From Strong's, Procedures in Experimental Physics, Prentice-Hall Inc., 1938)

25.10.2.2 Linfoot concludes that the variation of illumination, D(x', y'), of the knife edge is given, to a good approximation, by the equation

$$D(x', y') = \frac{1}{2\pi} \int_0^{\infty} du \int_{-\infty}^{+\infty} e^{(-iux' - ivy')}\, W(u, v)\, dv \tag{11}$$

Where

$u = \frac{2\pi x_1}{\lambda s}$

$v = \frac{2\pi y_1}{\lambda s}$

$i = \sqrt{-1}$

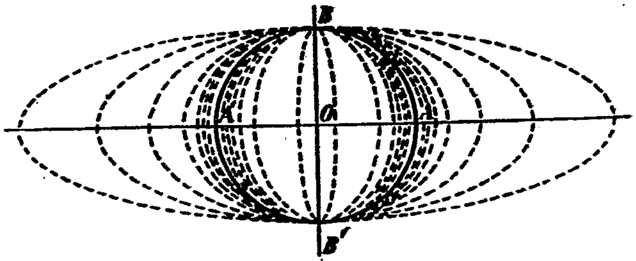
s = distance from the vertex of the surface to the knife edge.

$W(u, v) = \frac{1}{2\pi} \int \int E(x, y)\, e^{(iux + ivy)}\, dx\, dy$

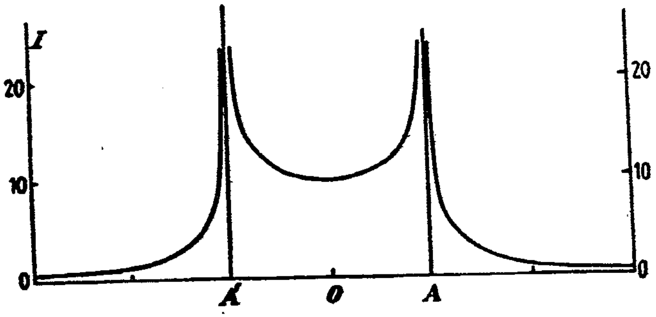
E(x, y) = illumination over the surface

Note carefully that (x, y) represent coordinates of a point on the surface under test; (x , y) represent the corresponding coordinates in the plane of the knife edge; and (x', y') represent the corresponding coordinates in the image of the x_1, y_1 plane.

25.10.2.3 An application of this equation to a true mirror with knife edge central gives Figure 25.27 (from Linfoot). Here the brilliant zone around the edge of the mirror is clearly predicted.



Isophotal lines for a true mirror under the Foucault test, with the knife edge central.



Intensity distribution along the horizontal diameter (after Linfoot).

Figure 25.27 - Isophotal lines and intensity distribution for a mirror under the Foucault test
(From Linfoot's, Recent Advances in Optics, Oxford Univ. Press, 1955)

25.10.2.4 If the knife edge is kept in the central plane but moved a distance, C, laterally we see the oscillations of Benerji (29). In this diagram $C' = \frac{2\pi C}{\lambda S}$, where C' is the distance moved in the plane of the image of the knife edge.

25.10.2.5 A practical problem frequently occurs in using the Foucault Test for surfaces of low reflection. The small pinhole that must be used results in very low light intensities, and the low light intensities make detection of the Foucault shadows difficult. To improve the situation a slit may be used. The slit must not only be narrow, but also of a length such that the aberrations of the system under study are effectively constant over this length. The details for interpreting the pattern resulting from this type of source are given in Linfoot (30).

25.10.3 The Foucault Test applied to non-spherical mirrors.

25.10.3.1 The Foucault Test can be used for paraboloidal as well as spherical mirrors. To simplify the interpretation, an additional flat is required as shown in Figure 25.28. For paraboloidal mirrors one may use several modifications of the basic Foucault test. One employs a flat mirror with a hole in the center. The arrangement is equivalent to that shown, but the observer looks along the axis of the surface being tested.

25.10.3.2 Another modification is the technique developed by Gaviola (31). This method is more sensitive than the basic test and is particularly useful as a guide in very close control of zonal errors. The experimental arrangement is shown in Figure 25.29. The Gaviola technique depends on the fact that for off-axis areas of a paraboloid the positions of best focus do not lie on the center line of the paraboloid but rather lie on a caustic which originates at the center of curvature. The method is essentially as follows. First the paraxial focal point is determined by the regular knife edge method. From this datum the equation of the caustic for the non-aberrated paraboloid is calculated. Next one calculates where the center of curvature (ξi, n_i) should be for a given facet or area. A knife edge is then used to determine where the center of curvature actually is - all of the paraboloid except the facet in question being covered up. The deviations $\Delta\xi$, Δn_i of the actual center of curvature from the ideal center of curvature for various facets are used to map the true surface of the mirror. Symmetry about the center line is assumed.

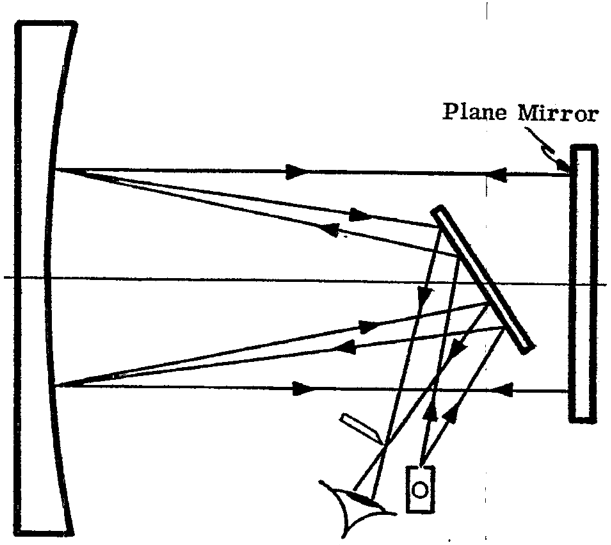


Figure 25.28- Foucault Test set-up for paraboloidal mirrors.

---

(29) loc. cit., (24)
(30) loc. cit., (28), 146
(31) Gaviola, JOSA 26, 163 (1936); also Strong, loc. cit., 23, p. 298

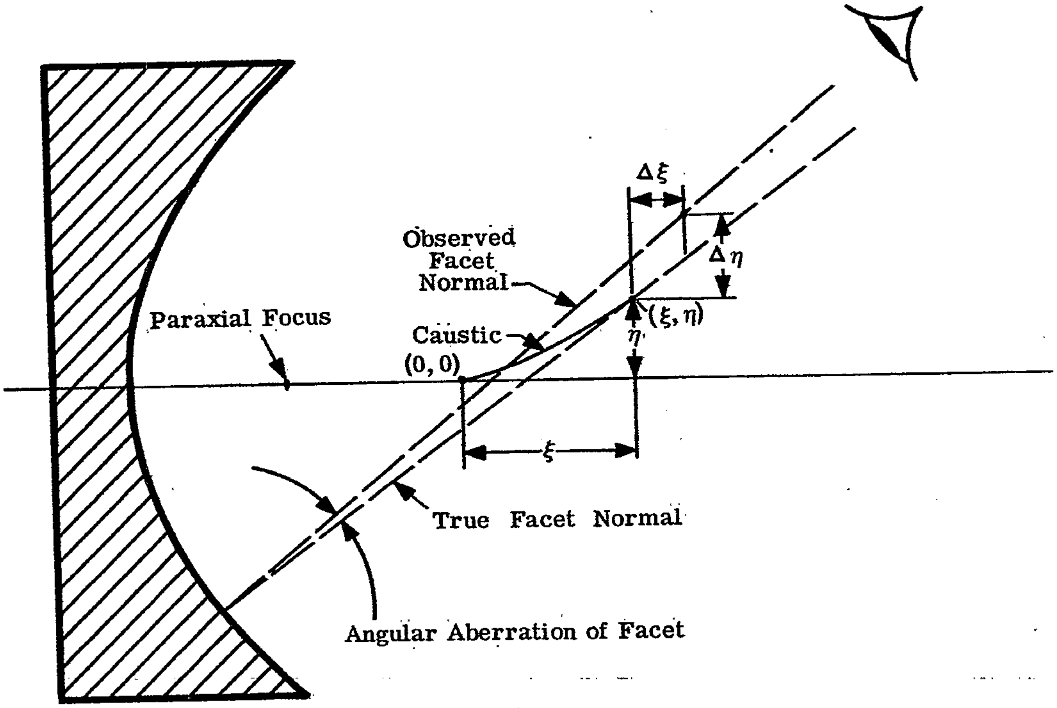


Figure 25.29-The Gaviola technique of Foucault Testing.

(After Strong's, Concepts of Classical Optics, W. H. Freeman and Co. 1958)

25.10.4 The Foucault Test applied to lenses. The previous discussion may have left the impression that the Foucault test is really applicable only to mirrors. This is not so. Basically any type of system may be tested with the same advantages accruing. Several examples are given in Strong (32). The essential technique is the same in each instance with modifications made as dictated by the system under test. In each case, it is the possibility of inspecting the zonal contributions more or less individually rather than seeing them in their integrated form which makes this test so important. Recently one of the leading precision optical makers commented that due to the increasing use of aspheric surfaces, virtually all of his lenses and mirrors were tested by this method.

25.11 THE STAR TEST

25.11.1 Introduction. It has been pointed out before that the choice of optical testing technique is frequently strongly related to the individual and his particular experience. It is generally acknowledged that H. Dennis Taylor carried the star test to its present heights and his disparaging comments on the knife-edge, or Foucault, test are interesting to read (33).

25.11.2 Technique.

25.11.2.1 The star test, as practiced by Taylor actually used a star as source of parallel light. The most frequently used star was Polaris although for checking achromatism, he also used the bluish star, Vega, and the reddish star Orionis. Today with the press of work, etc., it is seldom practical to depend on the visibility of the night sky, and artificial stars are used. One of the best "stars" is made by piercing a needle point, well honed, to varying degrees into layers of very thin tinfoil backed up by something like a very hard plastic or ebonite. Several trials should result in a very small perfectly round hole. The resulting aperture is then illuminated by a Pointolite lamp or equivalent and placed at the focus of a well corrected collimator as shown in Figure 25.30. Without much doubt, one of the best collections of star test photographs appears in Taylor's book and repeated also in Twyman's (34). It is reproduced in Figure 25.31. For convenience reference will be made using the figure numbers that appear in the photograph from Taylor, following in parentheses our figure number e.g. Figure 25.31(10a) is Taylor's Figure 10a, etc.

---

(32) Strong, Procedures in Experimental Physics 70-72 Prentice Hall (1953)
(33) Taylor, The Adjustment and Testing of Telescope Objectives, 50, Grubb, Parsons and Co. (1946)
(34) Twyman, Prism and Lens Making, 369, Hilger (1957)

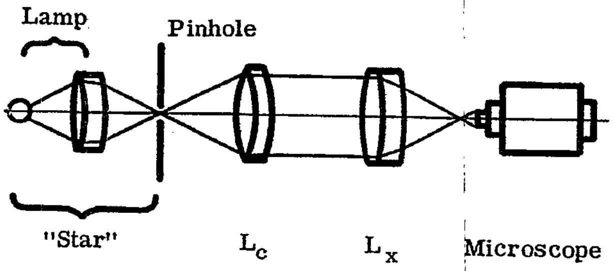


Figure 25.30- The experimental arrangement for a star test.

25.11.2.2 The star test technique consists of examining the image of the star with a fairly high power magnifier or telescope. It is pointed out clearly by Taylor that a careful examination of the observer's own eye is mandatory if a correct analysis of the system under test is to be obtained. The tests such as rotating the objective are simple to make. If the astigmatism rotates with it, the fault is with the objective, if not, the fault is with the eye.

25.11.2.3 In most instances more information is to be gained by examining the image out of focus and watching what happens as it goes through focus than trying to evaluate the system by an examination of only the in-focus image. The perfect figure will expand concentrically in an even fashion as the image passes through focus, with the intensity varying regularly in the ring structure. A perfect lens is shown in Figure 25.31 (17) while Figure 25.31 (18) shows the variation as a well-corrected lens passes through focus. It will be instructive to consider the principal star tests in the manner of Taylor, and this will now be done with frequent reference to Figure 25.31.

25.11.3 Squaring-on.

25.11.3.1 A telescope objective is considered "square-on" when the optical axis of the objective passes directly through the center of the axis or stated another way -- when the optical axis of the objective and the eyepiece coincide. Should the objective be cocked with respect to the eyepiece, the appearance of the image will depend upon the residual aberrations in the objective (assuming the eyepiece is effectively perfect). Usually the aberrations at focus that distort the image will be coma and astigmatism. Such a system is shown in Figure 25.31 (10a).

25.11.3.2 Reference should be made to Taylor for the actual process of squaring-on the objective, but the principle is clear from the photographs, viz. that the incorrectly adjusted objective will result in even the best focus being non-symmetric. Further, as the eyepiece is racked through focus, the image does not expand concentrically about the best focus image, but does so about a point to one side of the best focus image.

25.11.4 Achromatism.

25.11.4.1 It is a fundamental principle of instrumentation that we are interested in the performance of the whole instrumentation system and not just a part of it. Thus with visual optics and the known defects of the eye, we must design systems that take these defects into consideration. Occasionally we can put the defects to good use, but at all times we must be conscious of their effect upon the rest of the system. Even a perfect reflector will show a colored star test visually because of the achromatism of the eye. This may be checked easily and due account taken of it in judging systems whether they be reflective, refractive, or a combination thereof.

25.11.4.2 The defects of the eye are of course involved in the choice of eyepiece with which to judge the objective. Usually a fairly high power objective with a magnifying power of 50 - 100 times is suitable for use with a well-corrected eyepiece. Lower power objectives involve more of the eye aperture and consequently are affected more by the achromatism of the eye.

25.11.5 Astigmatism.

25.11.5.1 The nature of astigmatism has been discussed elsewhere in this manual so its details will not be reworked. It suffices one to say that the aberration known as astigmatism results in a star being focussed into two "lines" that are at right angles to each other, and displaced by an amount that depends upon the angle of view. This aberration is very easily detected with the star test. As the eyepiece is racked through focus one might see the image vary as in Figure 25.31 (12 d, d", d'). For corresponding positions inside and outside of focus one might see images as in Figures 25.31 (13) and (14).

25.11.5.2 A study of the photographs in Figure 25.31 and of the aberrational theory indicates that the star test is indeed a marvelously simple, and yet accurate, method for testing for astigmatism. It must be understood that seldom will a system have just one aberration and that particularly as one goes off axis, it becomes increasingly difficult to stipulate exactly the cause of the image degradation. It is here that experience plays such a vital role.

25.11.5.3 Assuming in the present case that only astigmatism is involved, one will usually find that the position of best focus will show a roughly circular image - the disk of least confusion - half way between the two astigmatic focal lines and of a diameter approximately equal to one-half the length of either focal line. While astigmatism is not as bad an aberration as some, if one is interested in "pointing" because of its symmetry, it may increase the spot diameter several hundred percent. This clearly decreases the resolution possible with the system. Once again we call the attention of the reader to the fact that the requirements of a good pointing system are not as stringent as those for a system of high resolution. This fact is too often overlooked.

25.11.5.4 As previously indicated, the defects of the eye must be taken into account and a truly stigmatic system may appear to the eye to be astigmatic. It is not only possible to separate the astigmatism of objective, eyepiece, and eye, but careful design can result in a system that shows no astigmatism to the eye, yet each component of the system, objective, eyepiece, and eye, each have demonstrable amounts of this aberration. Again the tests of the eye should be made initially with a low power eyepiece. More detailed drawings of star effects showing astigmatism, after Zernike and Nienhuis (35) are shown in Figure 25.32.

25.11.6 Zonal and marginal spherical aberration.

25.11.6.1 Perhaps the best way to get a feel for how the star test demonstrates zonal and marginal spherical aberration is to refer to Figure 25.33 where three possible extremes are depicted: (a) no spherical (b) marginal spherical and (c) zonal spherical. In each instance the lens under test (the element could of course be a mirror, etc.) is directed toward a distant star or equivalent as previously explained. In case (a) there is no spherical aberration at all and geometrically all rays come to point focus. Actually of course interference spreads the point into the familiar interference pattern and such a lens would show a perfect figure such as Figure 25.31 (17) at focus. Either side of, but not far from focus, such a lens might produce images such as Figure 25.31 (22) and (23). A glance at Figure 25.33 (a) demonstrates why this is so.

25.11.6.2 If the lens has marginal spherical aberration but little or no zonal, then we might see within focus an image similar to that in Figure 25.31 (15). Note carefully that there is no very bright center as contrasted with Figure 25.31 (22). Note also the way the intensity of the rings varies with transverse distance. Figure 25.31 (15) shows positive marginal spherical, i.e. the edge rays come to focus between the paraxial focus and the lens. Figure 25.31 (15a) shows negative marginal spherical aberration. The reason for Figure 25.31 (15) is again clear by reference to Figure 25.33 (b). Within focus there is a greater concentration of light for the marginal than for the central rays.

25.11.6.3 Where the marginal spherical aberration has been corrected, there may be residual zonal. This manifests itself, as might be expected, by an image of the form of that shown in Figure 25.31 (20) (inside focus) and 25.31 (20a) (outside focus). That there should be this high concentration in the third and fourth rings inside focus and in the 2 and 3 and 5th rings is reasonable providing the zonal error is as shown in Figure 25.33 (3c).

25.11.6.4 When checking for zonal or marginal spherical, the best technique is to check through focus and not just at focus. Further, the inspection should be made far enough from focus so that several rings appear as this is a more sensitive test. Generally speaking, the rules for interpreting zonal spherical aberration are

---

(35) loc. cit., (26), 96, Plate V

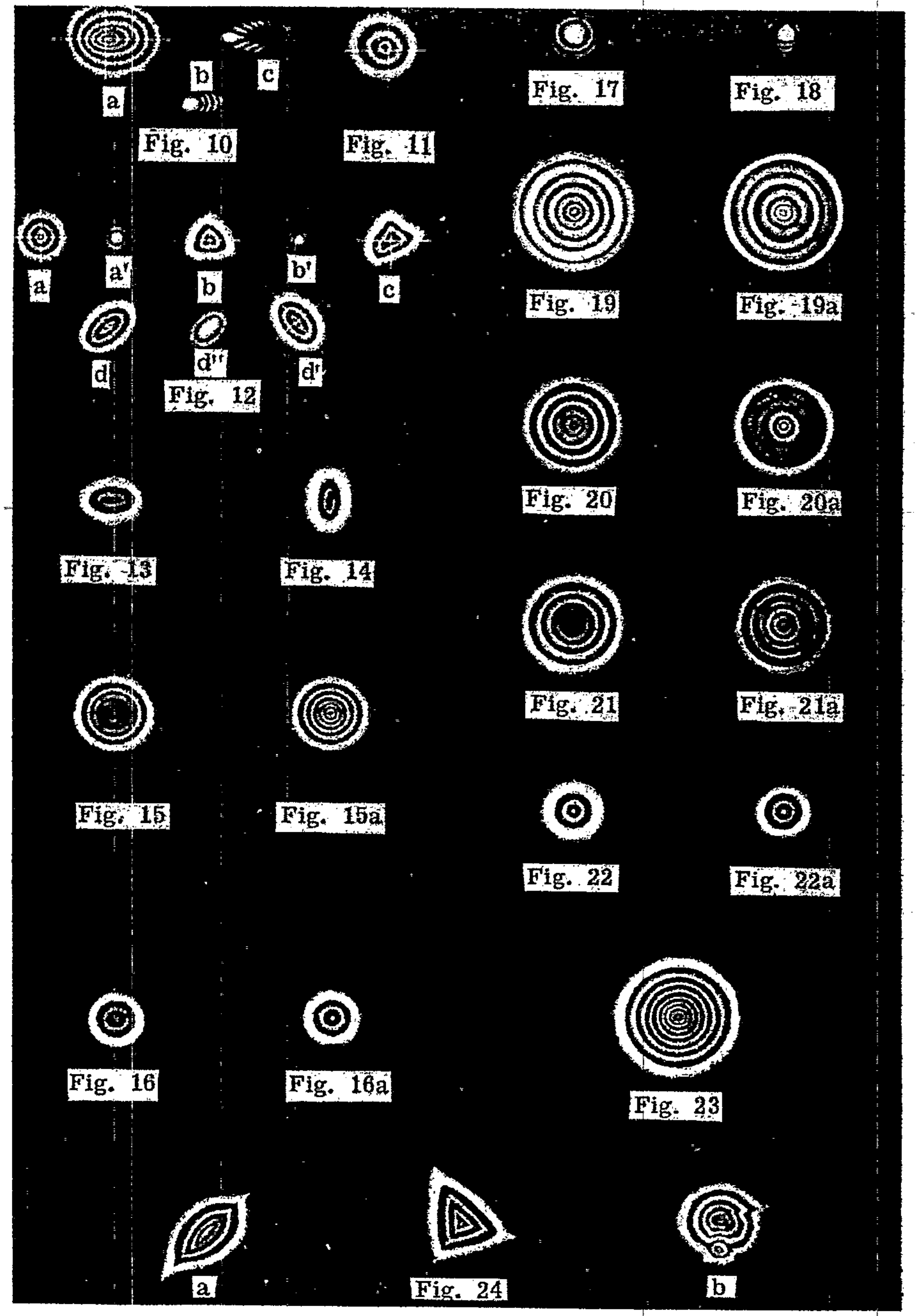
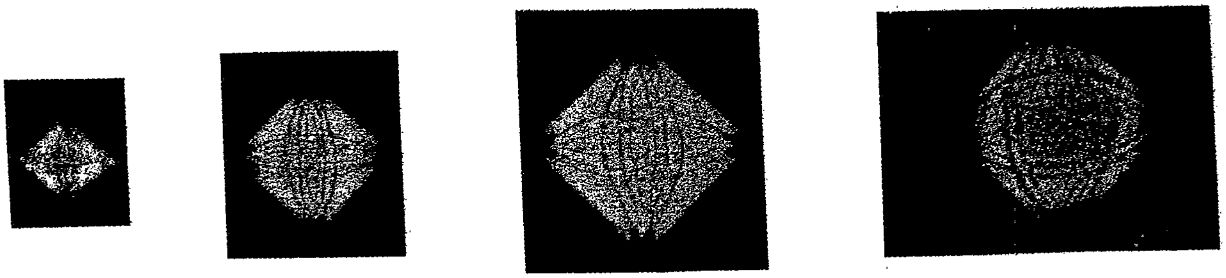


Figure 25.31-Star testing. (From Taylor's, The Adjustment and Testing of Telescopes Objectives, Grubb, Parsons and Co. 1946)

Fig. 10a. - Eccentric appearance of interference rings, due to the objective being out of adjustment.
c. - The focussed image of a star, when the maladjustment is about as much as in the last case.
b. - The focussed image, as visible when the objective is moderately out of square.
Fig. 11. - A section of the cone of rays taken closer to the focus, exhibiting a more moderate degree of maladjustment.
Fig. 12. - a, b, c, d, and d' are out-of-focus sections, as will be seen when the objective is correctly "squared on," and quite irrespective of other faults.
a', b' and d" are appearances of the focussed image corresponding respectively to a, b and d.
d, d and d" are also examples of astigmatism.
Fig. 13. - A section taken a very little way within focus, under a high power, exhibiting the fault of astigmatism.
Fig. 14. - The corresponding appearance to Fig. 13, as shown by a section taken at the same distance beyond focus.
Fig. 15. - Section within focus, showing result of positive spherical aberration.
Fig. 15a. - The corresponding section, taken at the same distance beyond focus.
Fig. 16. - A section taken closer to focus under a high power, exhibiting a slight residual spherical aberration; the central rings rather weak.
Fig. 16a. - The corresponding appearance at the same distance beyond focus; the central rings relatively strong.
Fig. 17. - The spurious disc or image of a star yielded by a perfect objective, and viewed under a very high magnifying power.
Fig. 18. - The spurious disc sometimes yielded by a large objective when resting upon three points, without intermediate supports being supplied to counteract the flexure due to the weight of the lenses.
Figs. 19 and 19a. - An example of marked zonal aberration, being sections of the cone of rays taken inside and outside of focus respectively.
Figs. 20 and 20a. - Another example of zonal aberration.
Figs. 21 and 21a. - Example of the general figure of an objective being tolerably good, but there is a region in the centre having a focus somewhat beyond the main focus.
Figs. 22 and 22a. - Two sections of the cone of rays yielded by a perfect objective, taken very near to and on opposite sides of focus, and viewed under a high power.
Fig. 23. - A section of the cone of rays yielded by a perfect objective, taken at about 1/4-inch on either side of focus, and viewed under a moderately high magnifying power.
Figs. 24 and 24a. - Examples of violent mechanical strain, due to imperfect mounting or bad annealing.
Fig. 24b. - Example of the effects due to the presence of veins in the material of the objective.

Index to Figure 25.31
(From Taylor's, The Adjustment and Testing of Teles. Objectives, Grubb, Parsons and Co., 1946)



± 1.4λ ± 2.7λ ± 3.5λ ± 6.5λ

Figure 25.32-Star tests showing astigmatism of varying degrees.

(From Linfoot's, Recent Advances in Optics, Oxford Univ. Press, 1955)

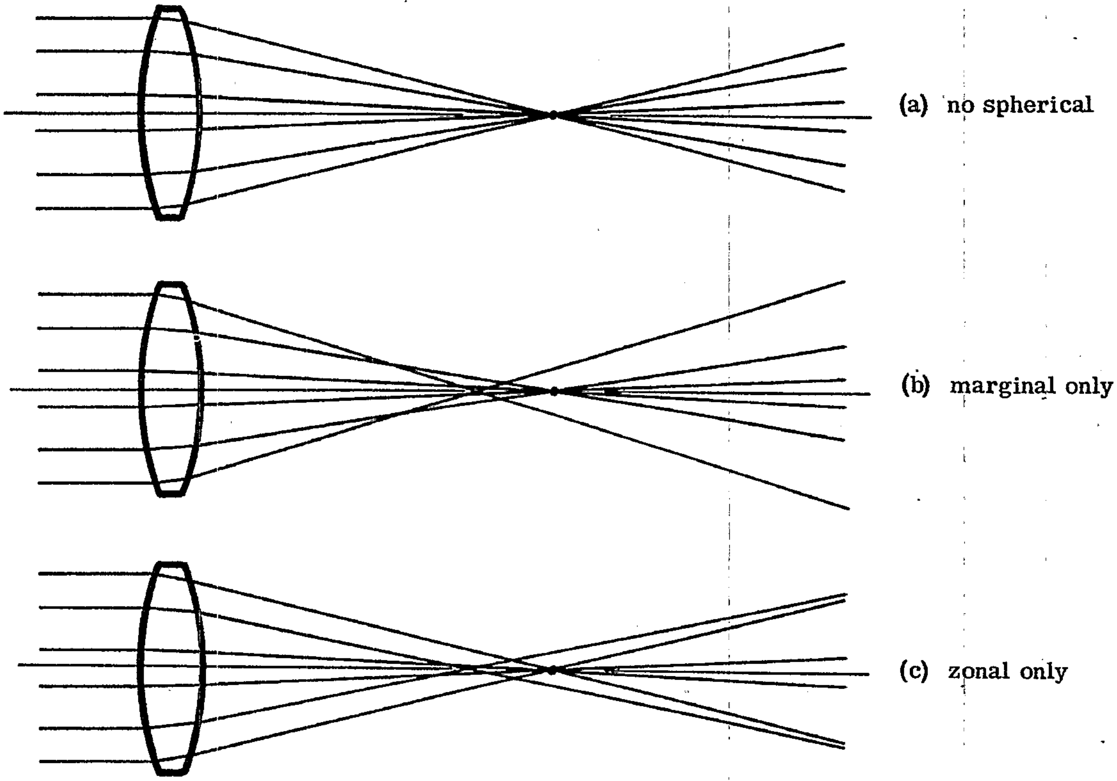


Figure 25.33- Various types of spherical aberration.

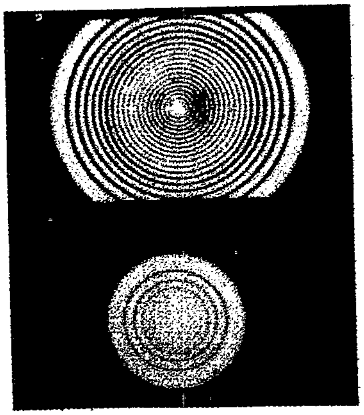
fairly straightforward with a bright zone or ring inside focus corresponding to a zone that focusses short. A bright zone or ring outside focus corresponds to a zone that focusses long. This corresponds of course to positive and negative zonal aberration. Again, Nienhuis (36) gives somewhat more detailed data, but not quite as much general information as Taylor. Figure 25.34 shows star images with varying degrees of primary spherical aberration at various focal positions.

25.11.6.5 It is clear that the star test furnishes a sensitive measure of the integrated effect of the whole lens. There is some question as to whether it gives as much information about a specific part of the lens or mirror as might a Foucault Test.

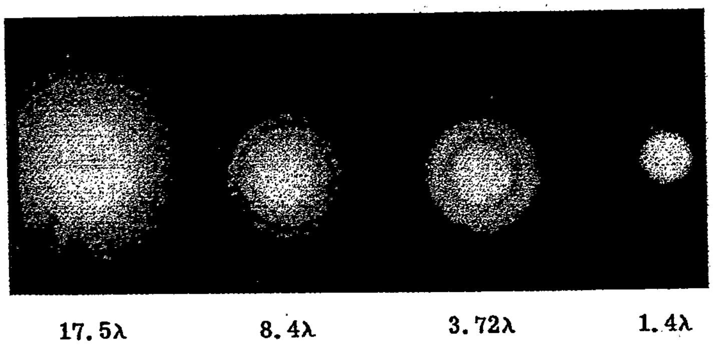
25.11.7 Coma. A good photograph of the effect of coma appears in Kingslake (37) and is reproduced in Figure 25.35.

---

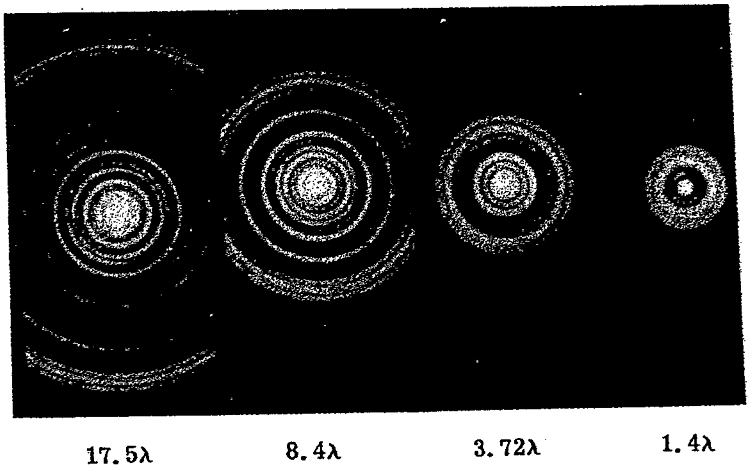
(36) loc. cit., (26) 48, Plate II; also, Thesis, Groningen (1948)
(37) loc. cit., (28) 84, Plate IV



(a) Images in presence of primary spherical aberration of amount 16λ, at marginal focus and at circle of least confusion.

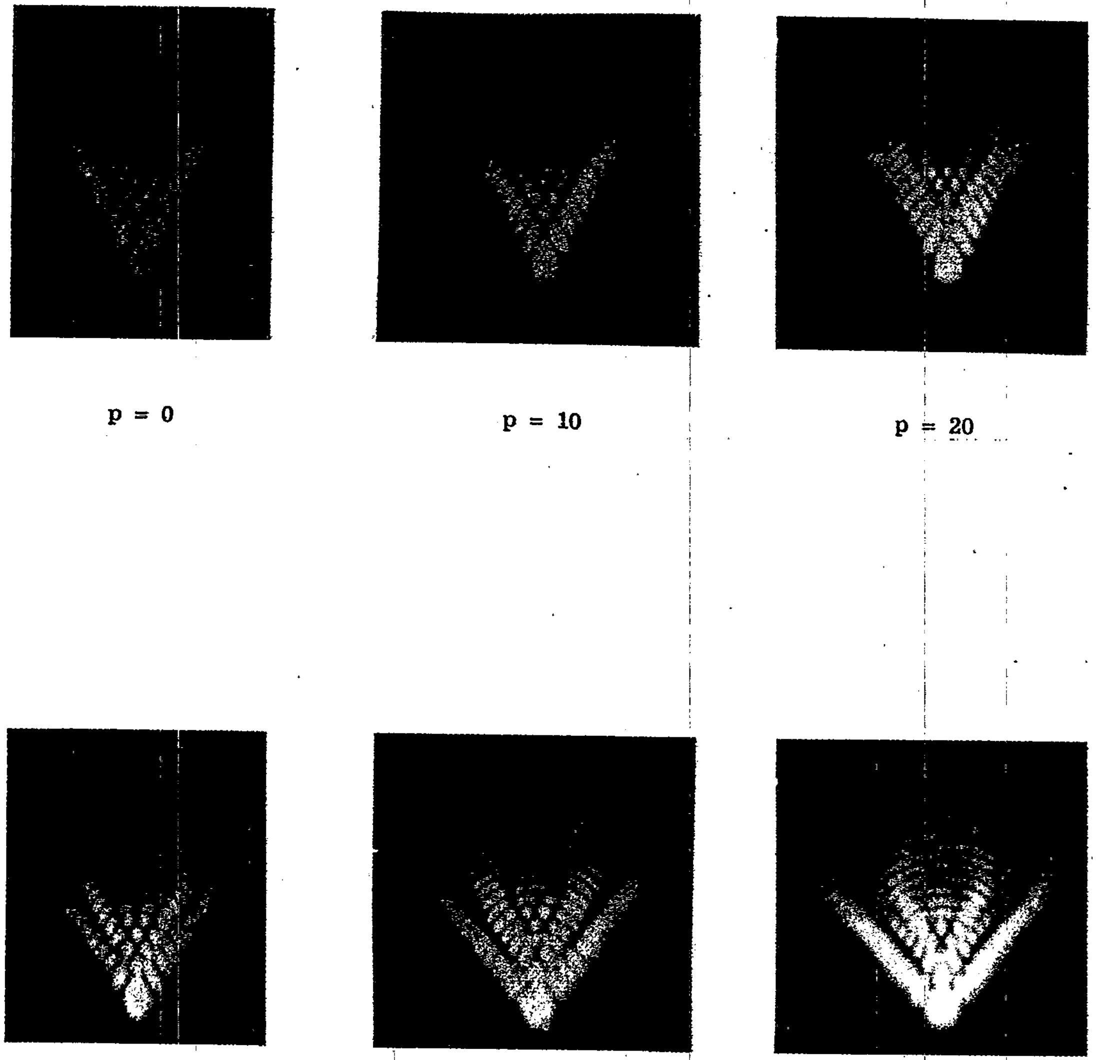


(b) Images in plane of paraxial focus, in presence of primary spherical aberration.



(c) Images in plane of least confusion, in presence of primary spherical aberration
scale three times that of (b).

Figure 25.34-Star images showing various amounts of primary spherical aberration.
(From Linfoot's, Recent Advances in Optics, Oxford Univ. Press, 1955)



p = 0

p = 10

p = 20

p = 30

p = 40

p = 50

Primary coma - ϕ is $2\lambda\,(r^3 - 2/3\,r)\cos x$ at focal settings corresponding approximately to the given values of p.

Figure 25.35-Star images exhibiting coma.
(From Linfoot's, Recent Advances in Optics, Oxford Univ. Press, 1955)

26 EVALUATION PHASE OPTICAL TESTS

26.1 RESOLVING POWER TESTS

26.1.1 Introduction.

26.1.1.1 The reason for the popularity of this general method stems from the feeling that, artistic considerations aside, the function of an optical system is to give information as to the detail in an object which is usually quite some distance away. Short of looking at the actual detail of the type on which the instrument under test is to be used, it has seemed reasonable to use some sort of artificial but definite target. Since many targets of military significance have sharp edges, targets with sharp edges seem to make sense. The nature of optical system performance is such that the edges should occur in at least two orientations and these preferably at right angles to each other. This deceptively simple process culminates then in a statement as to how many lines per millimeter can be resolved on the film of a camera, or as seen by the eye in a visual device. Actually it makes more sense to talk about a limit of resolution in terms of lines per unit solid angle, etc.

26.1.1.2 It will pay us to look somewhat more closely as to why this apparently straightforward process is called "deceptively simple". To begin, we have the fundamental question of what kind of target are we going to choose as a representative sampling of the in-use object. The USAF has been using the target in Figure 26.1 for years, while the National Research Council of Canada (1) has been using annuluses on a dark background as shown in Figure 26.2 along with a sector target proposed by Nutling. The U. S. National Bureau of Standards until recently used a line target as shown in Figure 26.3. This target and its applications were discussed in the reference cited. Recently NBS has adopted a new target and this is shown in Figure 26.4.

26.1.1.3 In addition to these, other groups have chosen targets made up of letters or numbers or combinations of special symbols or objects (2). To get informative as to the response of the optical system, at all angles, a target consisting of alternate black and white sectors has been used. (3) Apparently even the choice of the form of the target has been far from unanimous!

26.1.1.4 Let us look deeper. Even putting aside the question of form there is a considerable controversy over the contrast to be used between the dark and light portions. At least until the new NBS low contrast target came out, the British and Canadians were maintaining stoutly that the USAF high contrast targets were unrealistic as most of the objects photographed from an aircraft exhibited low contrast on the majority of days when photo-reconnaissance could be performed. We need not labor this point further except now we realize that not only the form but also the contrast is the subject of controversy.

26.1.1.5 With all this controversy the fact still stands that the system does have merit. Pestrecov (4) gave an excellent survey of the methods to date, and the serious student is referred to his work as regards relative merit of each. The particular claim for this technique is that it does give a single number that may be used to compare the performance of different lenses. The big question obviously is "granted it does give a figure by which to compare lenses but so do other techniques such as f- numbers, T- numbers, etc.", but does this really enable one to evaluate how a lens will perform in the field or does it merely tell how it would perform when photographing the very uninteresting lines and spaces on the test target? Unfortunately the answer to this question is not an unqualified "yes it does serve to state positively that this lens will be better than that in the field."

26.1.1.6 Having discussed these general ideas, let us now look at how the resolving power charts are actually used. As can be gathered from the above, different laboratories have their own techniques so we will sample three of the more common methods.

26.1.2 The NBS method.

26.1.2.1 Figure 26.4 shows the high and low contrast NBS charts. The dimensions of these patterns are given in the table below the charts. The contrast of the black on white is 1.4 while that of the black on grey is 0.20. The numbers on the chart "14, 20, 28", etc. refer to the number of lines/mm when this chart is used at a minification of 25X. The numbers refer to both the horizontal and vertical patterns whose linear extention

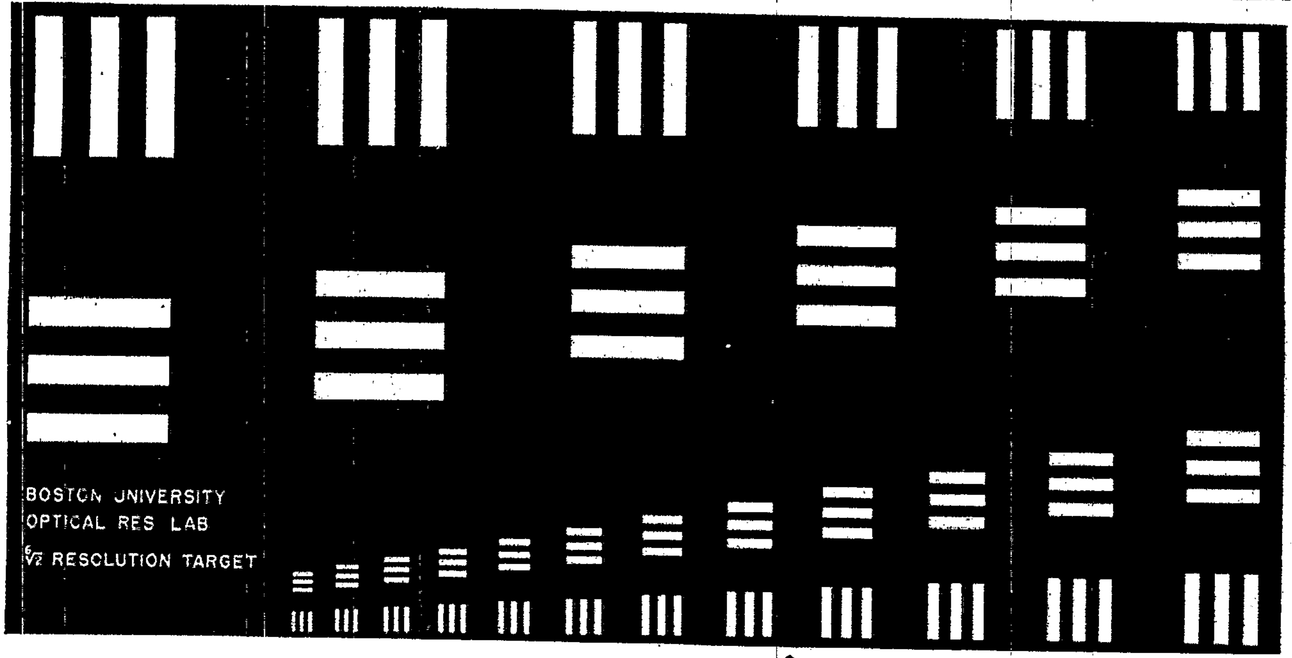
---

(1) Howlett, L. E., Photographic Resolving Power, Canadian Journal of Research, Vol. 24, Sec. A, No. 4, 15-40 (1946)

(2) MacDonald, NBS Circular 526, 51

(3) Jewell, A Chart Method of Testing Photographic Lenses, JOSA Vols. 2-3, Nos. 3-6, 52, (1919)

(4) Pestrecov, Photographic Resolution of Lenses, Photogrammetric Engineering, Vol. 13, (1947)

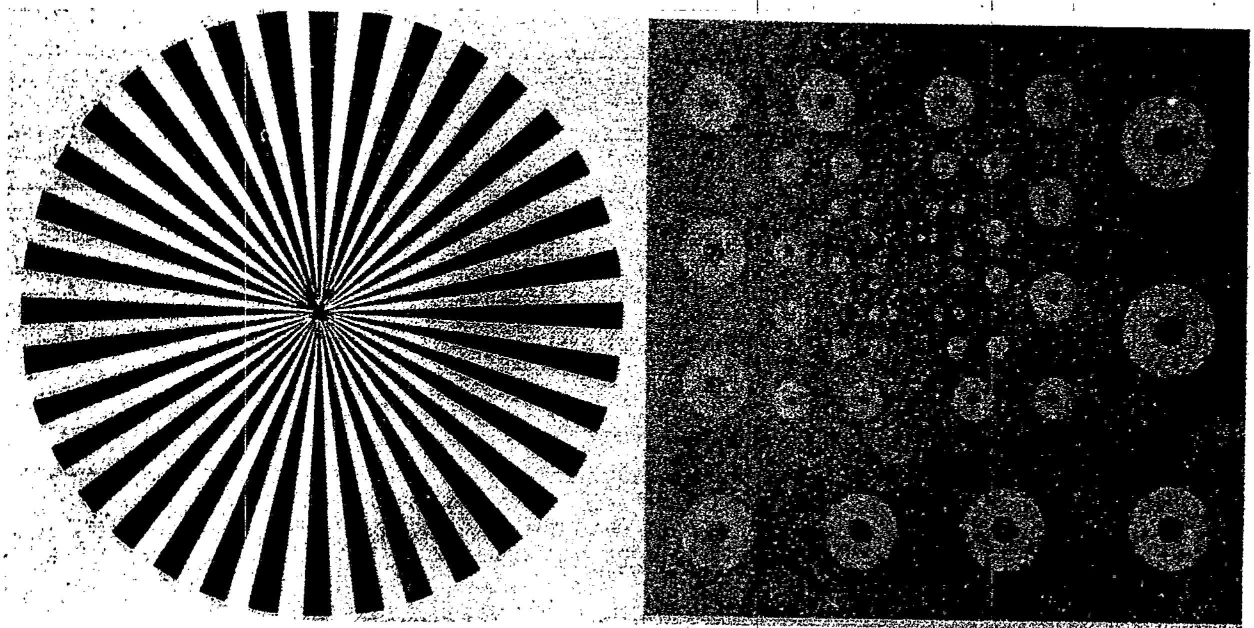


Calibration Sheet for B. U. O. R. L. $\sqrt[6]{2}$ Target

Unit	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Conversion Figure	21.5	24.1	27.1	30.4	34.2	38.3	43.0	48.3	54.2	60.8	68.3	76.7	86.0	96.6	108	122	137	153

These are resolution values for a B. U. $\sqrt[6]{2}$ target of 1mm. width. To determine resolution for each unit in lines/mm for any size target, divide each figure listed above by the width of the target measured from the extreme edge of unit 1 (the largest) to the extreme edge of unit 6.

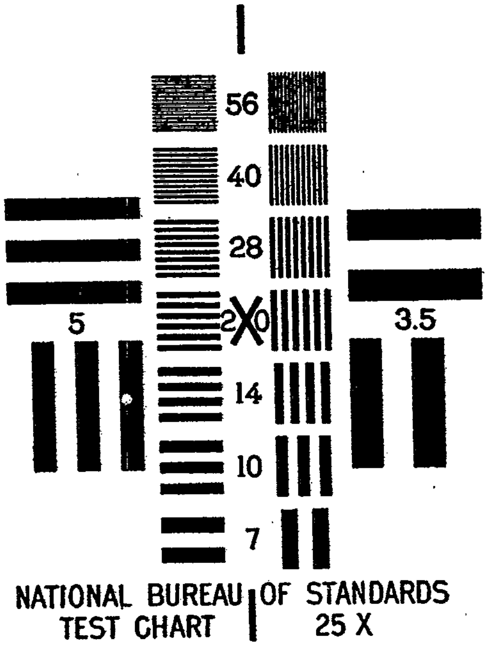
Figure 26.1 - The USAF resolving power target.



Sector target introduced by P. G. Nutting.

Canadian annulus target of 1.6:1 contrast ratio. The resolution values of the adjacent annuluses are in the $\sqrt[6]{2}$ ratio.

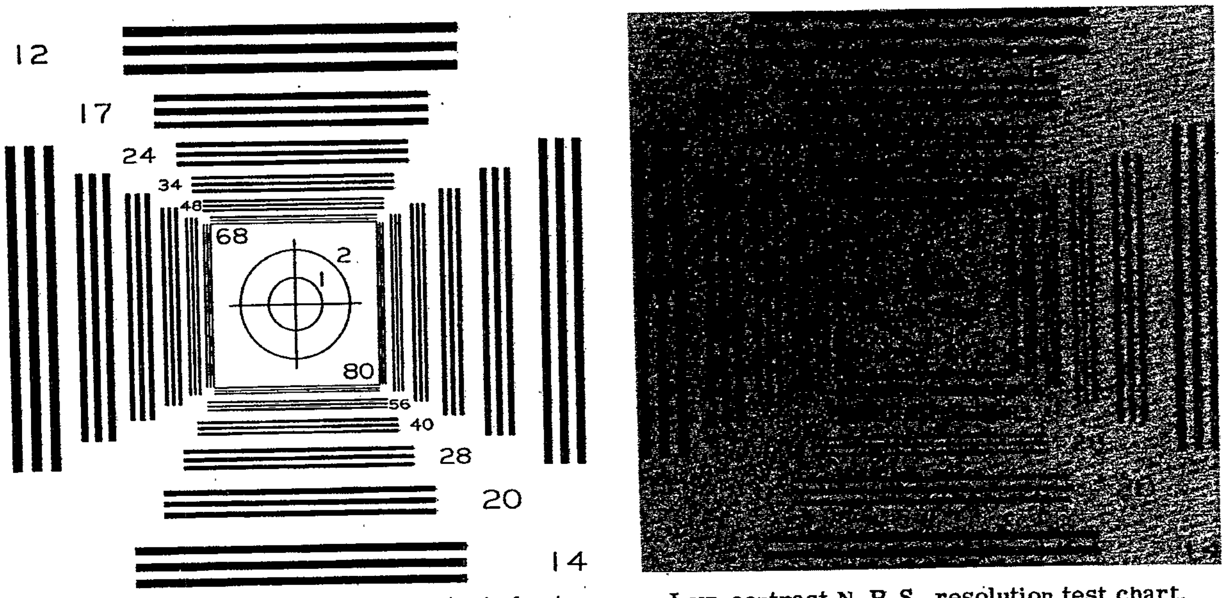
Figure 26.2 -The Nutting and Annuli resolving power targets.
(From Pestrecov's, Photographic Resolution of Lenses, Photogrammetric Engineering, Vol. 13, 1947)



This chart formed part of NBS Circular 428. The ratio of the line spacings in successive patterns of this chart is equal to $\sqrt{2}$. When the chart is photographed at the standard distance of 26f, the values of resolving power that can be measured with this chart range from 3.5 to 56 lines/mm.

Figure 26.3 - Old NBS resolving power target.

(A Test of Lens Resolution for the Photographer, NBS Circ. 428)



High-contrast N. B. S. resolution test chart

Low-contrast N. B. S. resolution test chart.

Pattern Number	80	56	40	28	20	14	68	48	34	24	17	12
Width of single black or white line	0.156	.233	.312	.446	.625	.893	0.184	.260	.368	.521	.735	1.042
Width of 3-line pattern	0.781	1.116	1.562	2.232	3.125	4.464	0.919	1.302	1.838	2.604	3.676	5.208
Width of space between patterns	0.781	1.116	1.562	2.232	3.125		0.582	.825	1.164	1.649	2.328	
Length of lines	18.0	19.6	21.9	25.1	29.6	36.1	18.0	19.6	21.9	25.1	29.6	36.0

Figure 26.4 -The new N. B. S. resolving power targets.

(Charts for Testing the Resolving Power of Photographic Lenses, F. E. Washer and I. C. Gardner, NBS Circ. 533(1953)) would run into the number. The chart used in this manner should be 26 focal lengths in front of the lens. The charts of course may be used both off as well as on axis. A common arrangement is to make a rack holding a series of the charts arranged in roughly the form of a square so that a photographic lens may be tested at all angles simultaneously. If the lens is to be tested visually, then it may be somewhat more desirable to reposition the test chart to the various angles of interest.

26.1.2.2 The observer after setting up the chart at the requisite distance determines which group is just distinguishable as three distinct lines and reports the corresponding number of lines/mm as the maximum resolving power of the lens at the given angle etc. Note that the measurement made in this way gives little or no information as to the response of the system to targets at fewer lines/mm.

26.1.2.3 Table 26.1 taken from NBS 533 shows the variation of resolving power of several hand held cameras. In this connection it is interesting to note the effect of using the high and low contrast targets. Inasmuch as we judge lines to be separated on the basis of contrast, it is important to note particularly Figure 26.5. The high contrast targets clearly may well be a more revealing as to what the actual resolution limits of the lens are. Further, the increased slope of the high contrast curve makes far more accurate measurements. Again we must warn that if the lens is to be actually used on low contrast targets, then we had better check it

Lens	EFL mm	F-number	Resolving power in lines per millimeter (angular separation from axis)									
			Tangential					Radial				
			0^o	5^o	10^o	15^o	20^o	0^o	5^o	10^o	15^o	20^o
A--	50	2	68	56	56	48	28	68	56	56	48	40
		2.8	68	68	68	68	56	68	68	68	68	56
		4	80	68	56	56	56	80	68	68	68	68
		5.6	80	80	68	68	80	80	80	80	80	80
		8	80	68	68	68	68	80	68	68	68	56
		11	80	80	80	80	68	80	80	80	80	80
		16	56	56	56	56	48	56	56	48	48	48
		22	56	56	48	48	48	56	56	48	48	40
B--	50	4.5	56	34	20	14	24	56	40	40	48	48
		5.6	56	28	17	20	34	56	40	40	56	56
		8	56	28	24	34	48	56	56	48	80	80
		11	56	34	34	34	56	56	56	56	80	80
		16	56	56	56	48	48	56	56	56	68	68
C--	85	2	68	68	34	17	--	68	68	48	34	--
		5.6	68	68	48	20	--	68	68	68	56	--
		11	68	68	48	24	--	68	68	68	80	--
D--	101	4.5	34	34	28	28	28	34	34	28	20	28
		5.6	40	34	28	28	28	40	40	28	14	28
		8	40	40	40	34	34	40	48	48	24	28
		11	40	48	48	40	40	40	48	48	34	34
		16	34	48	48	40	40	34	48	48	40	40
E--	101	4.5	28	28	24	12	7	28	34	34	28	20
		5.6	28	28	20	12	7	28	28	24	28	20
		8	34	28	24	17	14	34	34	34	28	28
		11	28	28	28	20	12	28	40	40	28	28
		16	34	34	28	17	12	34	40	40	28	20
		22	34	28	28	17	5	34	40	40	34	24
		32	34	28	24	17	12	34	34	34	34	28
F--			5.6	5.6	5.6	5.6	4.8	5.6	5.6	5.6	4.8	4.8

Table 26.1 - Resolving power at various apertures of several lenses of the type used on small hand-held cameras.

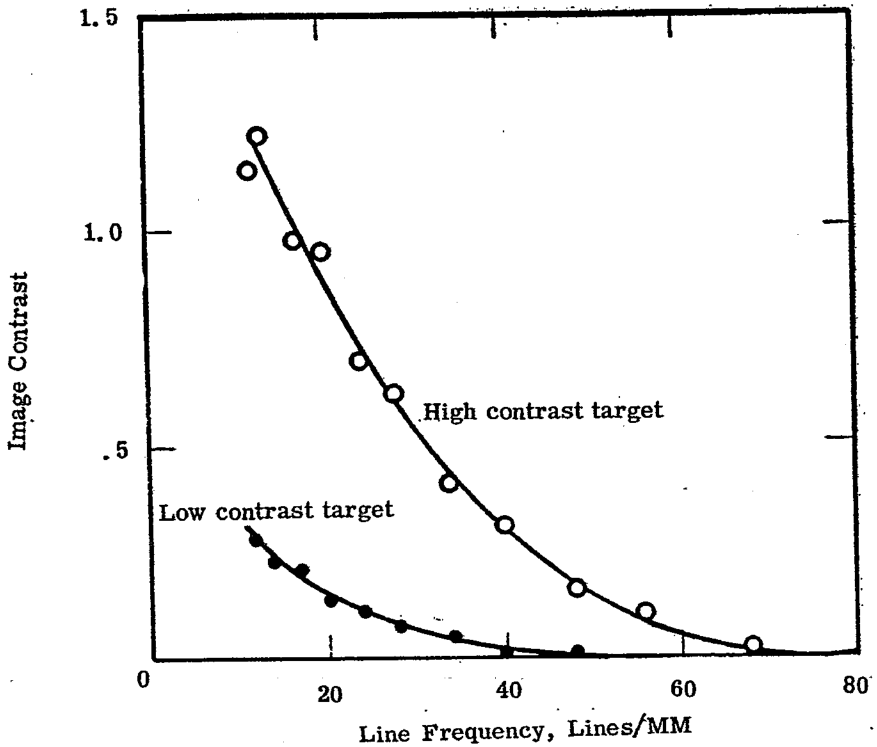


Figure 26.5-Variation of contrast in the image as a funtion of line frequency.
(Charts for Testing the Resolving Power of Photographic Lenses, F. E. Washer and I. C. Gardner, NBS Circ. 533(1953))

on low contrast targets. This also is shown clearly by Figure 26.5: if we want to resolve 50/mm at low contrast, then the lens examined is not suitable. If we want to resolve the same number of lines/mm at high contrast, then the lens might well be satisfactory. This is a crucial point in considering the usefulness of resolving power targets as evaluation tools.

26.1.2.4 Looking again at the question of visual optics such as binoculars, telescopes, periscopes, etc. we realize, as previously pointed out that here the most important characteristic is not lines/mm but rather lines/unit solid angle. We can also state this by saying that we are interested in the angular rather than the linear resolving power of the system. Tables 12 and 13 from NBS 533 enable the user to determine from the chart group just resolvable, the corresponding maximum angular resolving power for either the circles or the lines around them.

26.1.2.5 Care must be exercised in judging the resolving power of a visual system to be certain that the resolving power of the eye is taken into consideration. This means that the lines under study must all subtend an angle greater than that just resolvable by the eye--usually about 60 seconds of arc. This means then that the product of the resolving power of the target and the magnification of the system must be greater than, say 60 seconds of arc, if we are to obtain a true test of the resolving power of the system.

26.1.2.6 In this same connection, the resolving power of a sequence of optical systems is analogous to the effective bandwidth, or the effective rise time of a number of sequential amplifiers; the overall resolving power, Re , in terms of the resolving power of the individual components R_1, R_2 etc., is given approximately by

$$\frac{1}{Re} = \sqrt{\frac{1}{R_1^2} + \frac{1}{R_2^2} + \frac{1}{R_3^2} + \ldots} \qquad (1)$$

26.1.2.7 The NBS chart, when used in this standard manner, will cover a range of 14 to 80 lines/mm. For systems having higher or lower resolving powers the targets may be moved closer or further away. In some instances it may be convenient, where systems capable of resolving several hundred lines/mm are repeatedly encountered, to avoid the long working distances involved in the method above and reduce the targets photographically. Should this be done, great care must be exercised to see that the resolving power of the film and copying camera are such as to not degrade the targets.

26.1.2.8 While the NBS charts were developed primarily for lens studies, they may also be used as a basis for compliance with certain government specifications, for example

Federal Specification:

GGG-G-501b	Goggles, eyecup, protective, impact-resisting (chippers', grinders' etc.).
GGG-G-511a	Goggles, eyecup, protective (welders).
GG-T-621	Transits, 1-minute; and transit tripods.

Military Specification:

MIL-O-13830 Ord	Optical components for fire control instruments; general specification governing the manufacture assembly, and inspection of.

Commercial Standard:

CS159-49	Sun glass lenses made of ground and polished plate glass thereafter thermally curved.

26.1.3 The U.S.A.F. resolution target.

26.1.3.1 Originally suggested at the Bureau of Standards and carried to its present status by the U.S.A.F. Photographic Laboratory at Wright-Patterson Air Base, the U.S.A.F. target was designed primarily to evaluate the performance of aerial camera lenses. While the use of this target is controversial, it is probably the most widely used of all at the moment. The following comments of A. Katz (5), then of Wright Field, are much to the point. They were made during a discussion following a paper by R. E. Hopkins.

> "In connection with the points raised by Dr. Pestrecov and in earlier papers, I notice that a number of people have been gleefully trying to kick the three-line resolution target to death. I want to point out again--and I have done this in other meetings--that it has served its purpose well. This purpose, simply stated, is to serially grade lenses in a manner that will correlate with their photograph-making rank. I have yet to be shown that our use of the three-line target in the judging of lenses to be used for aerial photography has led to any error, let alone consistent error."
>
> "Now we have lots of data, most of which is not neat and packaged. The exigencies arising with the working conditions in the Air Force are such as to effectively preclude the careful running of planned experiments. We substitute large numbers of airplane flights and tests, and after a number of years we come to pretty definite conclusions--by statistical osmosis, if you will. We know by now that when we get a lens that performs well in the laboratory (on the much maligned three-line high-contrast target) it will take high-quality photographs in the air on good days as well as bad days. The converse is also true. Laboratory test enable us to predict the quality of actual aerial photographs. I can't expect much more of a laboratory test. Let us not forget that it is only within the last 10 years that lens performance began to be specified in terms of resolution requirement over the field and that manufacturers began to use these tests, and it is only within the last couple of years that photointerpreters have begun to hear of lines per millimeter as a measure of performance."

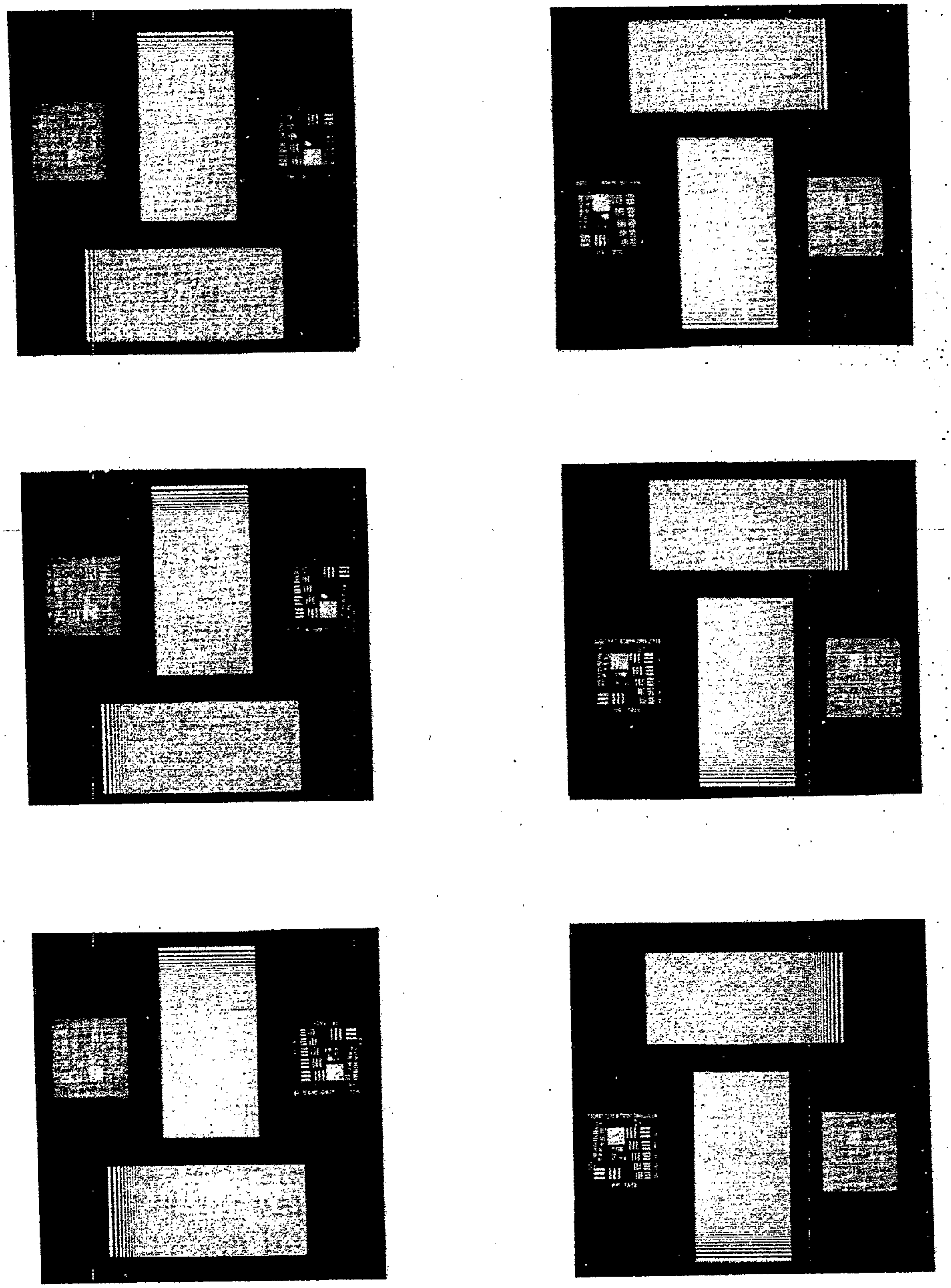
26.1.3.2 The type of tests (6) to which Mr. Katz is referring are well demonstrated in a portion of a series of through-focus trials shown in Figure 26.6 on a 40-inch f/5 Baker telephoto aerial camera looking through a window of poor quality. The target was the standard high contrast U.S.A.F. target plus a low contrast version of same plus a two variable frequency high contrast targets first introduced by Washer and Rosberry (7). The target was distant from the camera some 35 focal lengths to minimize the effects of spherical aberration. The term "window" here refers to the glass covering the hole in the skin of the aircraft through which the aerial camera sees the ground. More or less comparable focal positions are shown side by side for ease of

---

(5) NBS Circular No. 526, 200

(6) These tests were run by Mr. William C. Britton while at the Boston University Physical Research Laboratories and under a U.S.A.F. contract. Mr. Britton is now with Itek Corp.

(7) Washer and Rosberry, JOSA vol. 41, No. 9, 597, (1951)



window 45° Obliquity

No Window

Figure 26.6 - Resolution target testing for Baker 40" f/5 telephoto aerial camera.

comparison. The focal settings was changed by .005" between successive exposures. The variation of the resolution limit with the high and low contrast targets is clear. The effect of off-axis aberrations is also clear.

26.1.3.3 Composite target tests such as these demonstrate the difficulty of deciding on which target, if any, to settle on to the exclusion of all others. In fact it is pretty generally the opinion of the "conservatives" that no one target gives all the information that is needed to fully evaluate a lens. Were a given optical system always to be used on exactly the same type target, that would give a one to one correlation between laboratory testing and field performance. It, thus, is the very versatility of optical systems that gives rise to our difficulty.

26.1.4 The Kinetic Definition Chart.

26.1.4.1 There was developed during World War II (8) and subsequently improved upon (9 and 10) a routine system for checking the resolving power of visual optical systems. The system is essentially a resolving-power target approach, but incorporates many features not formed in the spur-of-the-moment setups commonly found in laboratories. The targets employed, as well as plan and side view schematics are shown in Figure 26.7.

26.1.4.2 The apparatus derives the word "kinetic" from the motorization of some of its parts, but the term is misleading nonetheless. A glance at the charts will show that they are of constant line spacing but of various contrasts and situated in four positions. The ratio of lines/spaces is 1:1, and is essentially the chart first advanced by Foucault (11) in 1858. The variation in line spacing required to determine the resolving power of a system is effected by the optical reduction unit. This unit consists of four highly corrected microscope objectives of focal length 4, 8, 16, and 32 mm. By varying the distance from the target to the reduction unit by the adjustable space gauge shown in Figure 26.7, the lines/inch may be changed from coarse to fine.

26.1.4.3 There are several interesting aspects to the KDC Apparatus. One of these is the "artificial sky" which not only simulates (by varying its illumination) the sky against which many objects must be seen, but also the stray light found in most optical systems. This apparatus thus takes into account not only the low control of the object itself, but also the surround so important in retinal response. Incorporated into the KDC Apparatus is a standard telescope with an aperture that is variable. This very carefully constructed telescope is of superior quality and allows the observer, in effect, to set up a standard against which the test instrument is compared. Once again we see a recognition of the need for removing as far as possible the limitations of the particular observer's eye from the testing procedure. Here this is done by inclusion of an auxiliary telescope of such magnification that the limit of resolving power is determined by the instrument under test rather than the eye. The rest of the system is rather straightforward and all designed to give maximum ease of assessment to the observer.

26.1.4.4 The final report on the NBS chart or the USAF chart is the resolving power limit of the system. In this technique the final report is called the K.D.C. efficiency and is defined as follows:

$$\text{KDC efficiency} = \frac{\alpha_e}{\alpha_i \, \text{Mi}} \times 100 \qquad (2)$$

where

α_i = minimum angle resolvable using the instrument under test.
α_e = minimum angle resolvable with the eye alone.
Mi = magnification of the instrument under test.

Clearly then, this definition is not a statement of the resolving power of the instrument alone, but rather it is a comparison of the effective improvement the instrument affords over the eye alone.

26.1.4.5 The factors directly preportional to α_e and α_i are conveniently determined directly from the KDC apparatus as follows. With the auxiliary telescope in place (if it will be required with the instrument under test as previously explained) the observer adjusts the target-to-turret spacing until the target is just resolved and the K.D.C. scale (lower left of drawing, just above the reversing switch) is read. The pointer on this scale is coupled to the target holder. The instrument under test is then inserted in its proper place and the K.D.C. scale again read. The K.D.C. efficiency is now obtained from the equation:

---

(8) NDRC Report (classified)
(9) Coleman and Harding, JOSA 37, 263, (1947)
(10) NBS, 526, 95, (1954)
(11) Foucault, Ann. de L'observation de Paris, 5, 197, (1859)

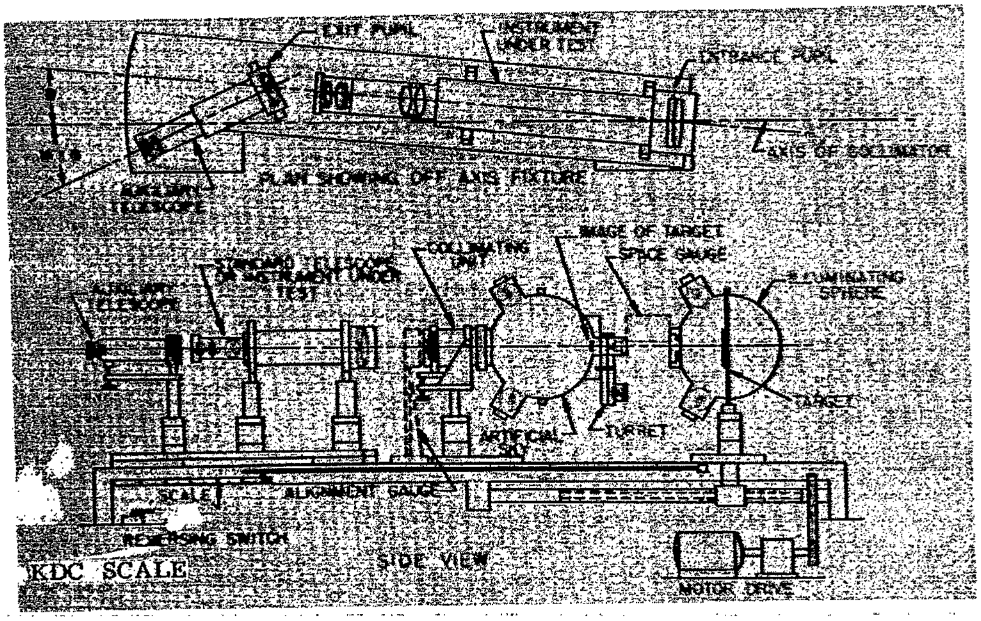
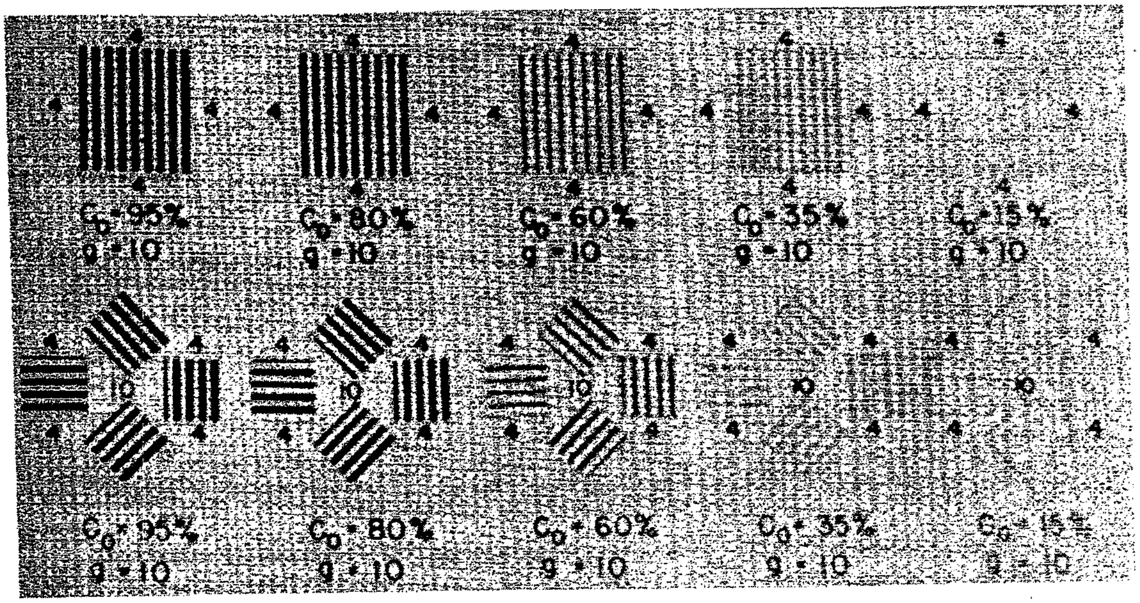


Figure 26.7 (a)- The KDC apparatus schematic.



Modified Foucault resolution targets.

$$C_o = \frac{B-b}{B} = \frac{\text{Reflectivity of white band} - \text{Reflectivity of dark band}}{\text{Reflectivity of white band}}$$

C_o = inherent target contrast: g = number of white bands per inch.

Figure 26.7 (b)- The KDC apparatus targets.

Figure 26.7 - The KDC apparatus and target charts.

$$\text{K.D.C. efficiency} = \frac{Xi}{MiXe} \times 100 \qquad (3)$$

where

Xi = K.D.C. scale reading with the instrument under test (and of course the eye)
Xe = K.D.C. scale reading with the eye alone
Mi = the magnification of the instrument under test.

If it is desired to compare a production instrument with the standard telescope, the K.D.C. reading taken with the standard telescope replaces Xe in the above equation.

26.1.4.6 There are many other uses of the K.D.C. apparatus but certainly its versatility and ease of manipulation recommend it when a large amount of work of this type must be done.

26.2 GENERAL DISCUSSION OF SINE WAVE TESTING

26.2.1 Introduction.

26.2.1.1 At about the time the controversy as to just what type of resolution target should be used was reaching its zenith, a paper given by Schade (12), an electrical engineer, brought to bear on the problem of optical system evaluation, the full resources of a completely different field viz., communication theory. While others such as Selwyn (13) and Duffieux (14) had preceded Schade in their investigations into this general area, there is little doubt in most minds that Schade (15) was responsible for focusing the attention of the optical world on the optical possibilities of this method.

26.2.1.2 It will be of interest to look briefly at Schade's original problem. Schade was studying the problem of optimizing the response of a television system starting with the optical pick-up in the studio through the electronic and electromagnetic systems to the final presentation on a kinescope in the home. His background here as an electrical engineer had taught him that one may study the response of an ordinary amplifier two ways (a) by feeding a single transient pulse to the amplifier and noting its response or (b) using sine waves of different frequencies and noting the phase shift and/or amplitude change as the sine wave signal passed through the amplifier. Fourier analysis shows that all the information contained in (b) is actually implicit in (a) but the transient is harder to use experimentally.

26.2.1.3 With the knowledge that this testing technique was a proven method, Schade in effect asked "why can't I do the same sort of thing for the optical part of the system? If I can do this, then I should be able to use the theories already developed for optimizing cascaded amplifiers." The question then arose as to what there was about an optical system that corresponded to the electrical sine waves. After the idea was conceived that the variation in intensity with angle as seen by the lens did indeed constitute a frequency, albeit a "spatial Fourier frequency" and not the frequency associated with $v = f\lambda$, the way was clear. There did remain then (and still does now) much theoretical work to do but at least the direction was indicated. The problem of translating the Fourier spatial frequencies into the temporal frequencies used in electronic amplifiers was easily solved by scanning techniques already under study in the sister field of flying spot scanner television.

26.2.2 Basic theory.

26.2.2.1 Inasmuch as this manual is not intended to develop all the pertinent theory but rather to acquaint the reader with possible methods, most of the details of the mathematical treatment will be omitted. The reader, however is invited to study closely the many excellent articles in this field. Some of these are in the following references: (16) through (23)

---

(12) Schade, A New System of Measuring and Specifying Image Definition: Symposium on Optical Image Evaluation, NBS, Oct., 1951. Proceedings published in NBS circuler 526, (1954).

(13) Selwyn, Theoretical Estimation of Combined Effects of Film and Lens on Resolution; RAE Report N. H. 698, April, (1940).

(14) Duffieux, L'intégrale de Fourier et ses Applications à L'optique, Besançon, Faculté des Sciences, (1946).

(15) Schade, Electro-Optical Characteristics of Television Systems, RCA Rev., 9:5-37, 245-286, 490-530, 653-686; (1948).

26.2.2.2 As indicated above and by Schade and Duffieux, an optical system may be considered as a two dimensional electrical filter. Further in electrical work we normally think in terms of amplitudes and at least in normal circuit work do indeed measure our signals by determining their amplitude. In optics, however, we cannot measure amplitude directly but instead measure intensity. A negative amplitude has no physical significance (although it can be interpreted as indicative of a 180° phase shift) for optics while it is a common and significant occurrence in electronics. As an aside we might note, however, that in the detection of electromagnetic radiation we can measure only power directly. The spatial frequencies to which we are referring are thus variations of intensity. This is an important point.

26.2.2.3 Let us assume that the coordinates in an object plane are denoted by ξ and η and in the image plane by x and y. The intensities in the object and image plane are then indicated by $O(\xi, \eta)$ and $i(x, y)$ respectively. We should note here that the terminology is not yet standardized and we are here following that of O'Neill (loc. cit. 16, p E-3). An object point $O(\xi, \eta)$ is then spread out into an image point $i(x, y)$, this "spread function" being denoted by $i(x, y)$. If we now apply this spread function to each point in the object, we can predict the appearance of the image by convolving the spread function with the object distribution according to equation (3).

$$i(x, y) = \int_{-\infty}^{+\infty}\int S(x-\xi, y-\eta)\ O(\xi, \eta)\ d\xi\ d\eta \qquad (4)$$

Assuming for the moment that this convolution is amenable to the techniques of the Fourier transform, we can do the same thing as (4) in the spatial frequency domain by utilizing equation (5).

$$i(\omega_x, \omega_y) = \tau(\omega_x, \omega_y)\ O(\omega_x, \omega_y) \qquad (5)$$

where $i(\omega_x, \omega_y)$ and $O(\omega_x, \omega_y)$ are the image and object expressed in terms of Fourier spatial frequencies and $\tau(\omega_x, \omega_y)$ is the so called "transfer function" of the system (for details see loc. cit. (16) p 232, et seq.)

26.2.2.4 Note clearly what has happened. We have replaced the convolution integral which is difficult to compute, by a product. The two equations of course say basically the same thing and their interrelationship is clearly seen by the more complete definition of the transfer function (loc. cit. (17), p26).

$$\tau(\omega_x, \omega_y) = \frac{1}{2\pi} \int_{-\infty}^{+\infty}\int S(x, y) e^{i(\omega_x x + \omega_y y)}\ dx\ dy \qquad (6)$$

Clearly we must be able to calculate, or otherwise determine, the spread function in (4) before the $i(\omega\ ,\ \omega\ )$ may be calculated theoretically. This is a sizeable task. It turns out, however, to be relatively simple to do it experimentally and this is effectively where the art stands at present. The technique, based on experimental determinations of $\tau(\omega_x, \omega_y)$ has lead to a new, although still controversial, method of evaluation of optical systems. Synthesis by use of this principle as a design method is still in its infancy.

26.2.2.5 Let us back off again and look at why equation (5) is so important an evaluation tool. The reason rests in part on the fact that the object and image are related in the spatial frequency domain by a multiplicative factor while in the spatial domain they are related by a complex summation. If we have two systems one of which clearly shows a better high frequency response than the other, we can be sure that this system will have the higher resolving power. Further the process of obtaining the sine wave response, or $\tau(\omega_x, \omega_y)$, will give (or usually does) the response at all frequencies and not only at the maximum resolvable condition as with the resolution target system. It is thus, theoretically, possible having $\tau(\omega_x, \omega_y)$ to predict the image for any object by the use of equation (5).

---

(16) Proceedings of Symposium on Communication Theory and Antenna Design AFCRC - TR-57-105 (ASTIA Document No. AD117067). While this symposium was aimed primarily at antenna designers, the organization of it was such that not only is the optics covered rather well by O'Neil and Parrent but also the basic mathematics and physical requirements are outlined in detail. One should note particularly the bibliography prepared by Parrent on Page M-1.

(17) O'Neil, Selected Topics in Optics and Communication Theory Itek Corp. (1958) Note - This publication has an exceptionally complete bibliography of work in this field.

(18) O'Neil, Publications of the Theoretical Optics Section, Itek Corp. (1958)

(19) Marechal, The Contrast of Optical Images and the Influence of Optical Aberrations, NBS Circular No. 526, p9, (1954)

(20) Elias, Optics and Communications Theory, JOSA, 43, 229, (1953)

(21) Hopkins, H. H., The Frequency Response of a Defocussed Optics System, Proc. Ray Soc (London), 321A, 91, (1955)

(22) Blanc-Lapierre, Upon Some Analogues Between Optics and Information Theory, Symposium on Microwave Optics, McGill University, (1953) - Proceedings published by Antenna Section Air Force Cambridge Research Center.

(23) Parrent and Drane, The Effect of Defocussing and Third Order Spherical Aberration on the Transfer Function of a Two Dimensional Optical System, Optica Acta, 3: (1956)

26.2.2.6 As might be expected, nothing is ever quite this rosy. Always there is the needle in the haystack or thorn in the rose. The difficulty here lies in the fact that the transformation from equation (4) to equation (5) presupposes that the optical system is perfectly linear and invariant over the object and image fields. Unfortunately this does not hold very well in poor systems. In good systems Linfoot and Fellgelt (24) have shown, however, that over the normal working field the assumptions are reasonably valid. A rather good discussion of the restrictions involved in making the jump from (4) to (5) has been given by Zucker (25) both for the case at hand, optical systems, and also for the allied problem-antennas. Much as it would be interesting to go into here more of the basic theory, the limitations of space require that we get on to the actual experimental techniques of measuring the transfer function and its applications. The interested reader will find the references given, however, replete with pertinent information. There are several methods of determining $\tau(\omega_x, \omega_y)$ or the equivalent, of which the following are representative only.

26.3 SINE WAVE TESTING WITH SINE WAVE TARGETS

26.3.1 The Schade system.

26.3.1.1 In Schade's original presentation, he demonstrated a system that, stripped to its basic features, was essentially that shown in Figure 26.8 wherein F represents a continuous film with a series of discrete sine wave targets. Each target was made by varying the intensity of the exciter lamp in a sound track camera sinusoidally with time while the film was moving at a constant rate through the camera. Sections of the film are shown in Figure 26.9 (26). P is a projector that allows the test pattern to be seen at any effective distance from the system under test, S. The light from S is focussed (usually with the aid of an auxiliary microscope) onto a scanning aperture, A. This aperture might be of any shape but usually it is most convenient to use a circle. Behind the aperture is a photomultiplier tube, PM, which feeds into a recorder, R.

26.3.1.2 In action then, the film moves through the projector producing a spatial frequency sine wave. The fact that the film is moving means that there will be a sinusoidally varying electrical signal from the photomultiplier tube. The sine wave response is then given simply by the ratio of this ac signal at a spatial frequency, N, to that which the system would give if the frequency were extrapolated to zero. In Schade's terminology $r_{\bar{\psi}} = \bar{\psi}_n/\bar{\psi}_o$ where $r_{\bar{\psi}}$ is the sine wave response. Typical sine wave response factor curves are shown in Figure 26.10. These response curves were taken from research done in this field by Shack (27) when at the National Bureau of Standards. Figure 26.11 from the NBS Report gives the variation of $r_{\bar{\psi}}$ with focal position for a fixed spatial frequency while Figure 26.12 gives the variation of $r_{\bar{\psi}}$ with focal position for a fixed color. Figures 26.13 and 26.14 show the variation of $r_{\bar{\psi}}$ with spatial frequency for different colors. Note the negative amplitude in these figures. It is due to a 180° phase change. Schack's apparatus was much the same as Schade's but Shack used a scanning slit instead of a scanning pinhole.

26.3.2 The Lamberts system.

26.3.2.1 Lamberts (28) and Lamberts, Higgins, and Wolfe (29) have studied the sine wave response particularly in connection with their lens evaluation program at Eastman Kodak. The reader will find Lamberts' article particularly interesting as he not only discribes the basic theory very lucidly but also presents a rather novel variation on the fundamental method.

26.3.2.2 In the Schade method the scanning aperture is very small and usually circular or square. In the Lamberts system the scanning aperture is a long slit. By the use of the slit it is possible to replace a target whose intensity varies sinusoidally by a target with a variable area as shown in Figure 26.15. This type of target has also been used by Lindberg (30). The scanning slit is indicated by SS in Figure 26.16. It can be shown the light distribution in the image is given by,

$$F(x) = b_o + b_1 \; |A^*| \; \cos\,(2\pi\gamma x - \phi) \qquad (7)$$

Where b_o and b_1 have the meaning shown in Figure 26.15, and b_1/b_o is the "normalized amplitude" as discussed in Lamberts' article. γ is of course the spatial frequency and x is the shift of any particular aspect

---

(24) Linfoot and Fellgelt, On the Assessment of Optical Images, Trans. Roy. Soc. (London) 247, (1955)

(25) Zucker, loc. cit. 5, p L-1

(26) Schade loc. cit. 1, p 233

(27) Shack, Investigations Into the Correlation Between Photographic and Photoelectric Image Evaluation, NBS Report No. 5483

(28) Lamberts, JOSA 48, 490 (1958)

(29) Lamberts, Higgins, and Wolfe, JOSA 48, 487 (1958)

(30) Lindberg, Optica Acta 1, 80 (1954)

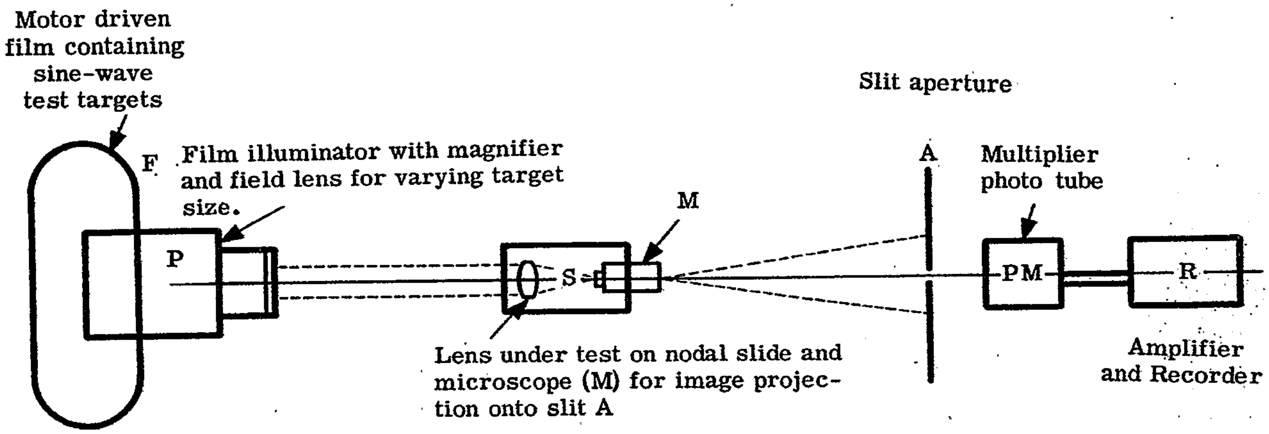

Motor driven film containing sine-wave test targets
F
Film illuminator with magnifier and field lens for varying target size.
P
M
S
Slit aperture
A
Multiplier photo tube
PM
R
Lens under test on nodal slide and microscope (M) for image projection onto slit A
Amplifier and Recorder

Figure 26.8 - The basic Schade system for determining the sine wave response of an optical system.
(Based on O.H. Schade's, Electro-Optical Characteristics of Television Systems, RCA Review, Vol. 9, 1948)

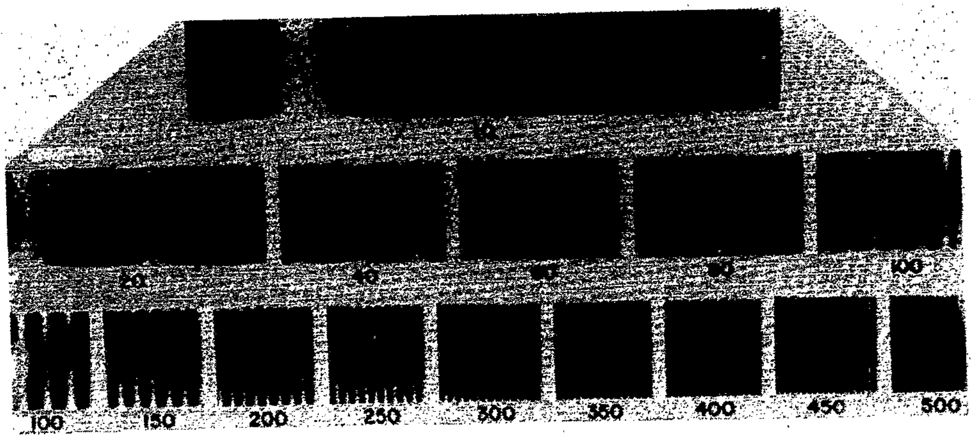


Figure 26.9 - Sine wave test targets.
(Based on O. H. Schade's, Electro-Optical Characteristics of Television Systems, RCA Review, Vol. 9, 1948)

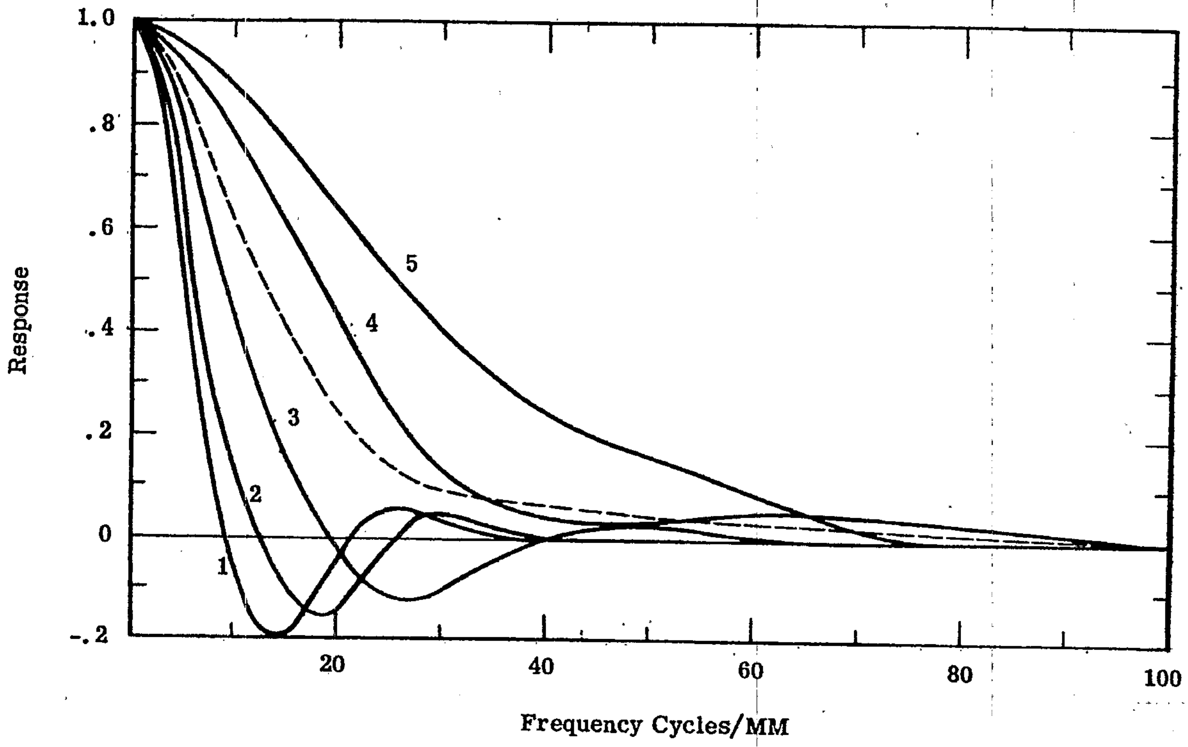


Figure 26.10 - Sine wave response factor vs line number (frequency) of lens A, .4 mm inside focus. In this and in following figures, curves numbered 1, 2, 3, 4 and 5 were obtained with Wratten filters 29, 25, 90, 16 + 60, and 45 respectively. The dashed curve was obtained with no filter. (Extracted from National Bureau of Standards Report No. 5483, Investigations into the Correlation between Photographic and Photoelectric Image Evaluation, R. Shack, under Air Force Contract Number 33(616)56-16)

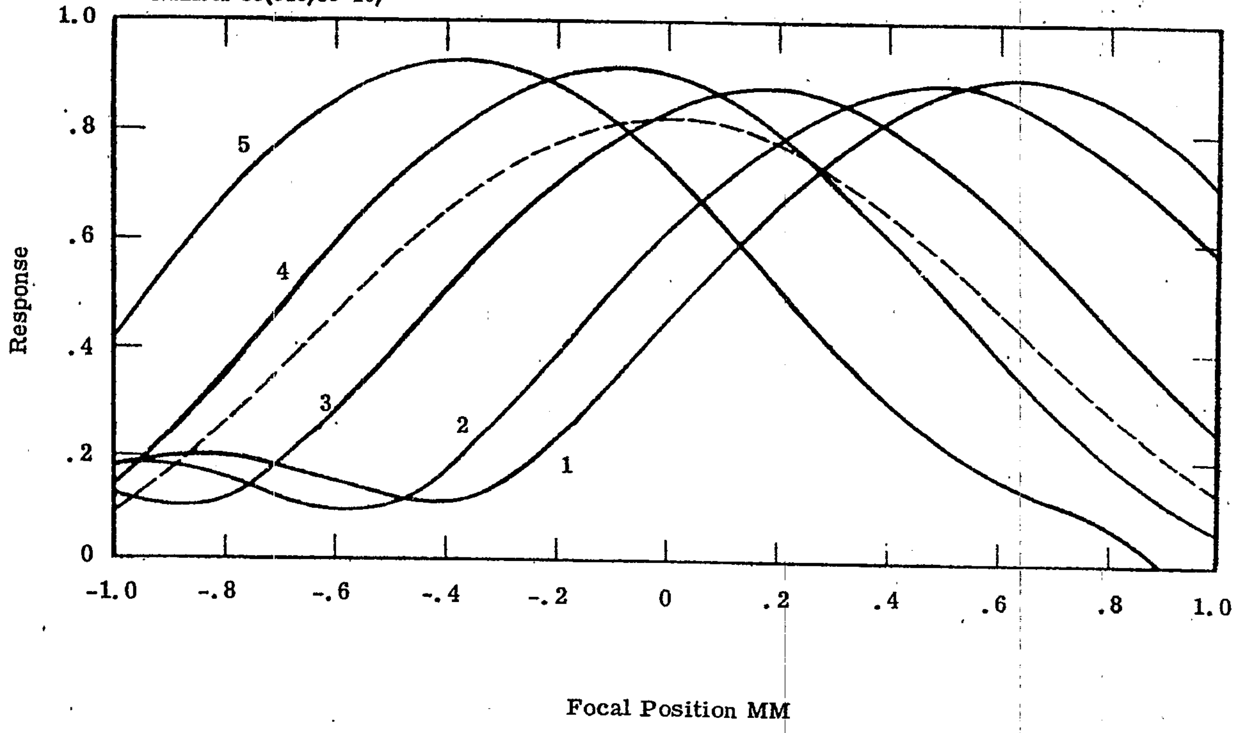


Figure 26.11 - Variation in response with focal position for lens A for different colors at a fixed frequency. The frequency chosen was 8 cycles per mm. (Extracted from National Bureau of Standards Report No. 5483, Investigations into the Correlation between Photographic and Photoelectric Image Evaluation, R. Shack, under Air Force Contract Number 33(616)56-16)

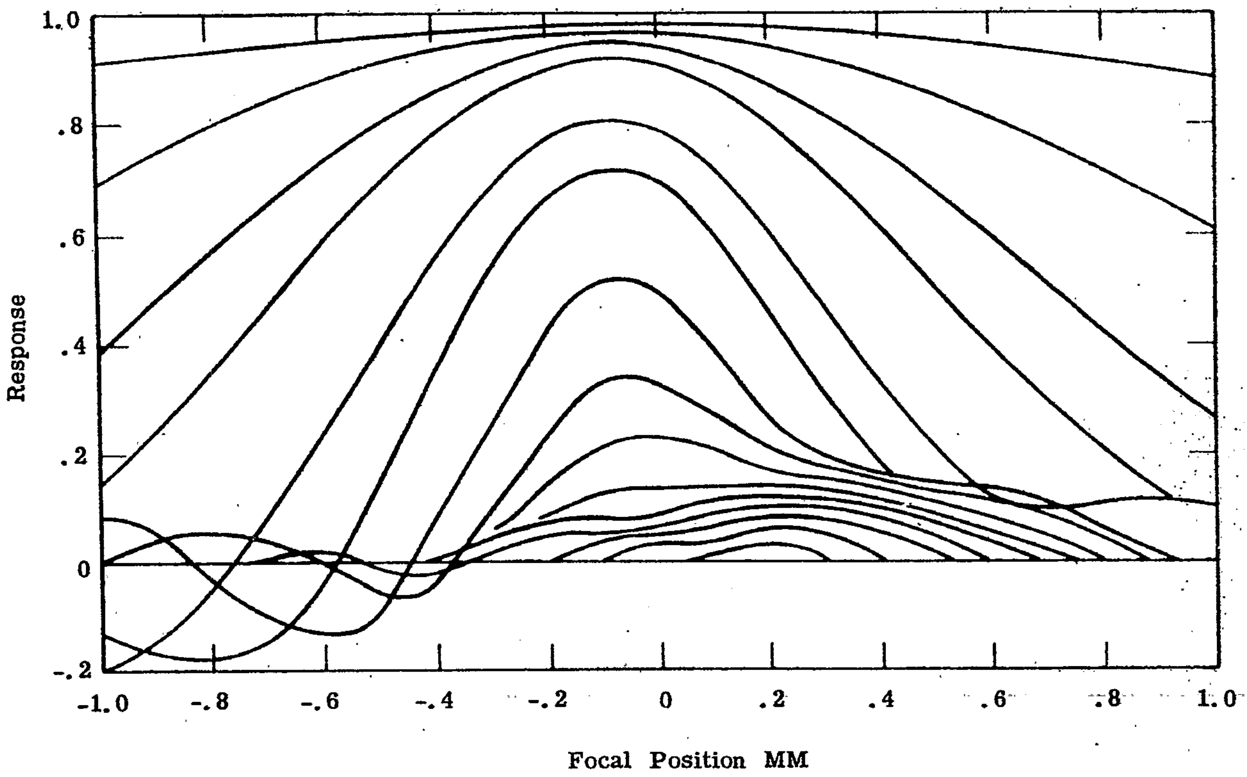


Figure 26.12- Through-focus response curves for lens A with filter 16 + 60.
(Extracted from National Bureau of Standards Report No. 5483, Investigations into the Correlation between Photographic and Photoelectric Image Evaluation, R. Shack, under Air Force Contract Number 33(616)56-16)

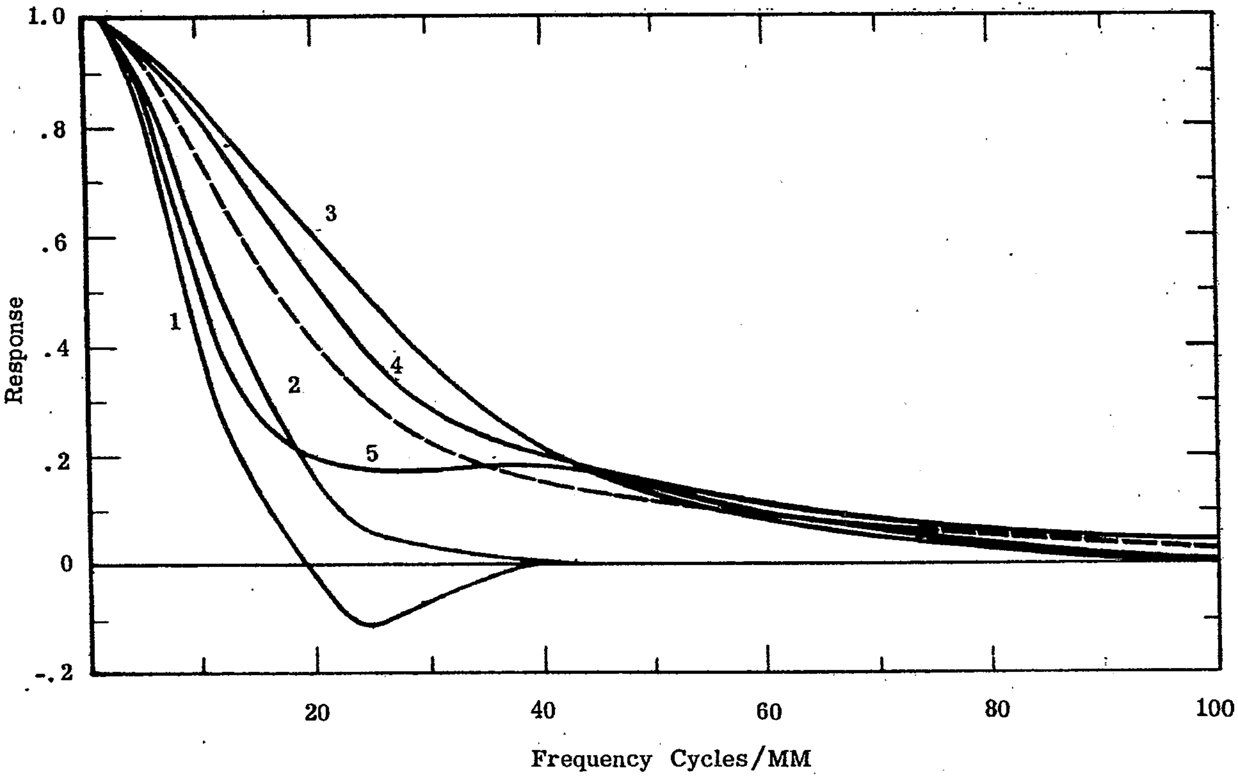


Figure 26.13- Frequency response of lens A at focus for various colors.
(Extracted from National Bureau of Standards Report No. 5483, Investigations into the Correlation between Photographic and Photoelectric Image Evaluation, R. Shack, under Air Force Contract Number 33(616)56-16)

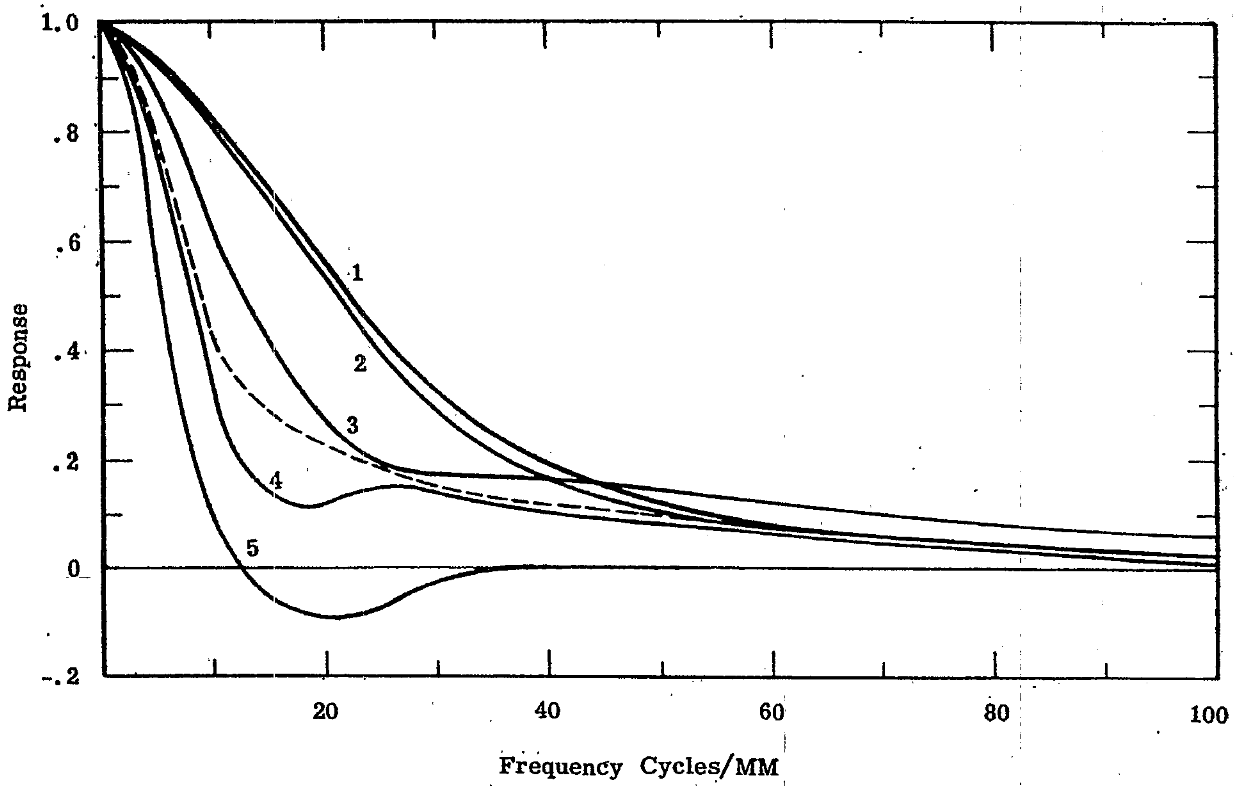


Figure 26.14- Frequency response of lens A .4 mm outside focus for various colors. (Extracted from National Bureau of Standards Report No. 5483, Investigations into the Correlation between Photographic and Photoelectric Image Evaluation, R. Shack, under Air Force Contract Number 33(616)56-16)

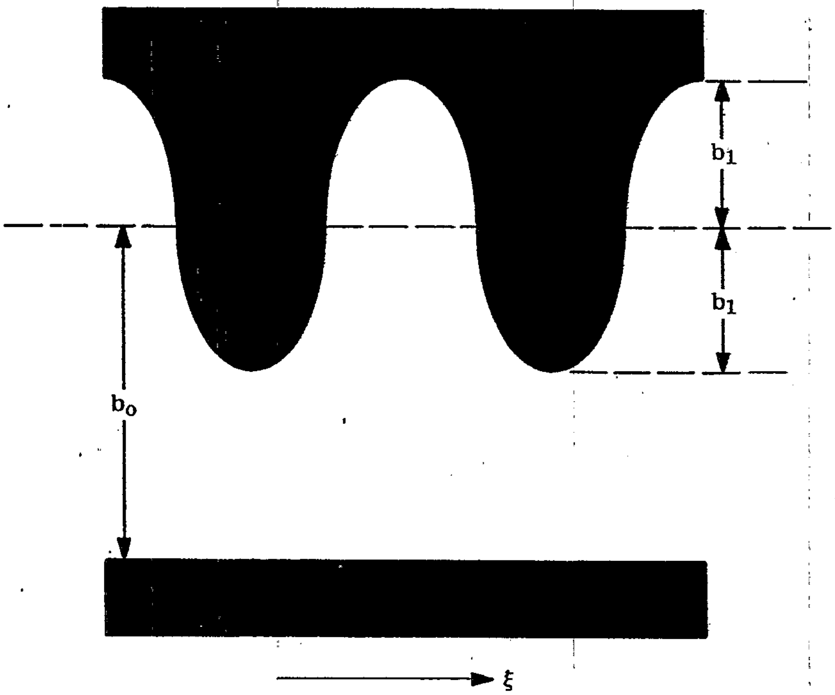


Figure 26.15- The Lambert's test object for measuring sine wave response. (Extracted from National Bureau of Standards Report No. 5483, Investigations into the Correlation between Photographic and Photoelectric Image Evaluation, R. Shack, under Air Force Contract Number 33(616)56-16)

of the target. ϕ is the spatial phase angle between object and image and is something not covered specifically in Schade's original work. A* is the sine wave response previously defined. The lens bench used in the experiments set up to confirm the theory is shown schematically in Figure 26.16. TO is the test object, which for this work was either a slit (used to determine the spread function) or the target shown in Figure 26.9. L is the lens under test, with SS being the scanning slit, and P the photomultiplier and recorder ensemble. T is a tangent bar arranged to tilt the object and lens, L, when studying off-axis response.

26.3.2.3 The action of the system is similar to that of the basic method, and the reader may refer to the original article for further details. The reader should pay particular attention to the excellent discussion of the significance of the spatial phase angle, and the symmetric and asymmetric spread functions. Attention is also called to the discussion of the derivation of the spread function from the sine wave response. This is important when one remembers that in the introduction to this section the spread function was defined first, with the sine wave response introduced subsequently as a dependent variable. The fact that the one may be calculated from an experimental determination of the other bears out the statement made earlier about their relationship.

26.3.2.4 The significance of phase angle is pointed up in discussions of objects that represent coherent, incoherent, or partially coherent sources. Even with the simple systems checked by Lamberts the phase angle was a strong function of spatial frequency. Figure 26.17 shows both the normalized amplitude in percent (directly relatable to sine wave response via equation 7) and the phase angle as a function of spatial frequency in lines/mm for a certain lens.

26.3.2.5 Stephens (31) has recently indicated an interesting way of determining experimentally not only the cosine of the phase angle but also the sine of the phase angle. The advantage is that of increased precision for angles up to 45°.

26.3.3 The recording electronic lens bench of Herriott.

26.3.3.1 The recording lens bench we are about to describe is a long way from the first exploratory efforts in this field. Actually this lens bench is similar in purpose to the K.D.C. apparatus in that each was designed not so much to do research work as to check out large number of lenses routinely by their respective techniques. The target for this apparatus was first made by W. Herriott (32) and is shown in Figure 26.18. Note carefully that the spatial frequency varies continuously on the actual target with samples taken discontinuously along the length of the film to show the variation in the spatial frequency. The scanning slit is oriented vertically with respect this page. The target is on a 36 in. strip of 35mm film with 50 parallel opague tracks on 0.010 in. centers. The slit is a few microns wide and long enough to span most of the width of the 50 tracks.

26.3.3.2 In use the target film is wound around a drum inside of which is the light source and appropriate motors and clutches. Attention is called to the fact that the target does not directly present a sinusoidal variation of intensity to the optical system under test. The scanning slit, however, integrates the image over its length and the result is effectively the same as with the Schade system. The complete schematic layout of the system is shown in Figure 26.19.

26.3.3.3 In this method the sine wave response is measured by the contrast rendition which is defined as $\dfrac{\text{"image max - image min"}}{\text{object max - object min}}$. Defined in this way, the result is independent of the contrast in the object, a point about which there was much discussion in connection with resolving power targets. The contrast rendition is plotted automatically as a function of spatial frequency. A typical recording showing the result of a through focus test is shown in Figure 26.20. (33)

26.4 SINE WAVE TESTING WITH SQUARE WAVE TARGETS

26.4.1 General discussion.

26.4.1.1 One of the problems involved in sine wave testing is the actual production of the sine wave targets themselves. This has proved to be a major problem, particularly so as the demands of the theorists got tighter and tighter. One method has already been outlined above. Other techniques have been developed (34 - 36) but the fact remains that it is still easier to make a square wave target than a sine wave target. The question has naturally arisen "can we not utilize the known Fourier sine wave content of a square wave to produce the equivalent of a pure multiple frequency sine wave target?" The answer is "yes" with some restrictions. If

---

(31) Stephens, Computation of Achromatic Objectives, NBS, (1954)

(32) Herriott, W., JOSA 37, 472 (1947)

(33) Herriott, D., JOSA 48, 968 (1958)

(34) Kapany and Pike, JOSA, 46, 867 (1956)

(35) Kapany, Eyer, and Shannon, JOSA 47, 103 (1957)

(36) Kelly, Lynch, and Ross, JOSA 48, 858, (1958)

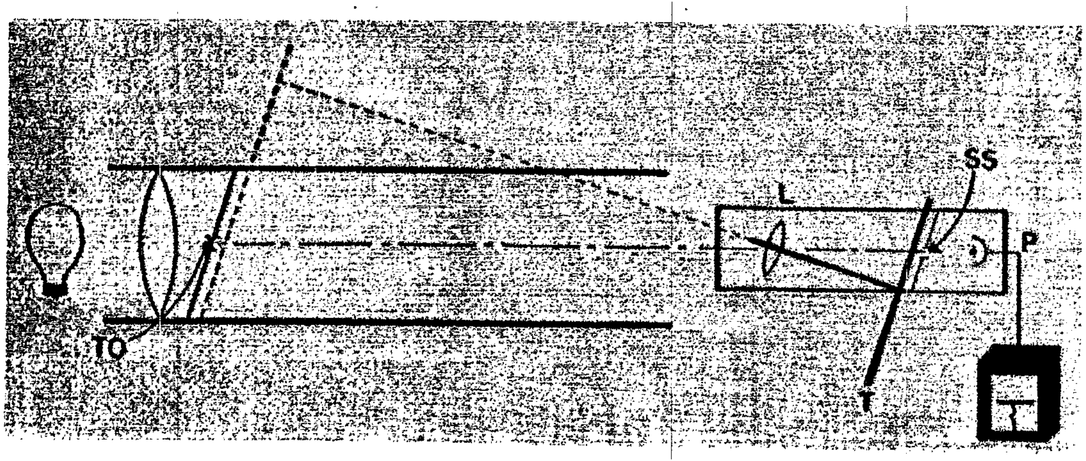
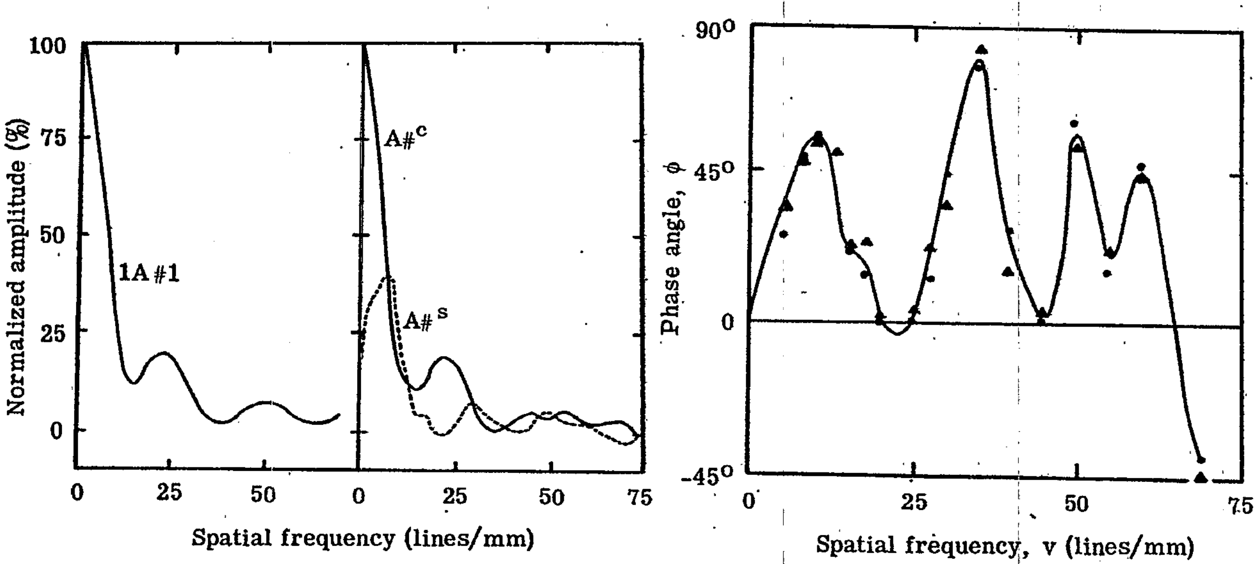


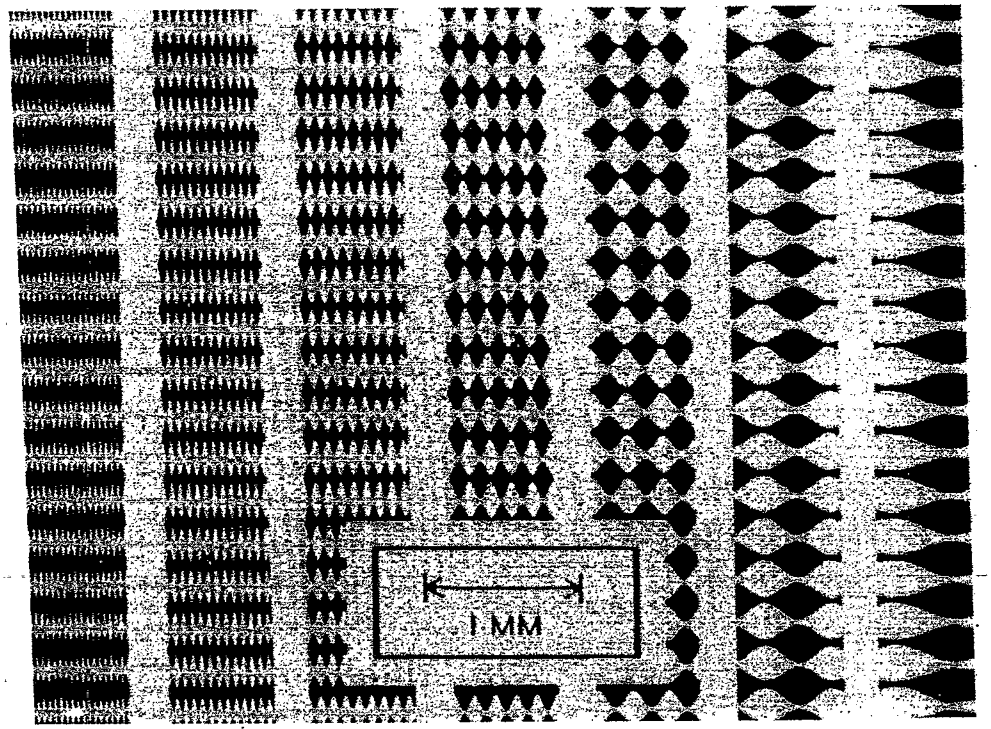
Figure 26.16 - Lambert's lens bench for determining sine wave response factors.



(a) Sine-wave response |A#| (left) and Fourier transfoms of it A#^{c} and A#^{s} (right) for a certain lens. A single sinusoidal test object was used to obtain |A#| and a double test object for A#^{c}; A#^{s} was computed from the other two.

(b) Phase angle as a function of frequency for the lens of fig. (a). The curve represents the mean of the two determinations ● and ▲.

Figure 26.17 - Normalized amplitude and phase angle as a function of spatial frequency.
(From Jour. Optical Soc. America, Lamberts 89, 1958)



Enlarged photographs at intervals along a sinusoidal target on which the frequency change is continuous.

Figure 26.18 - The Herriott continuous spatial frequency target for determining sine wave response. (From Jour. Optical Soc. America, W. Herriott 37; 472, 1947)

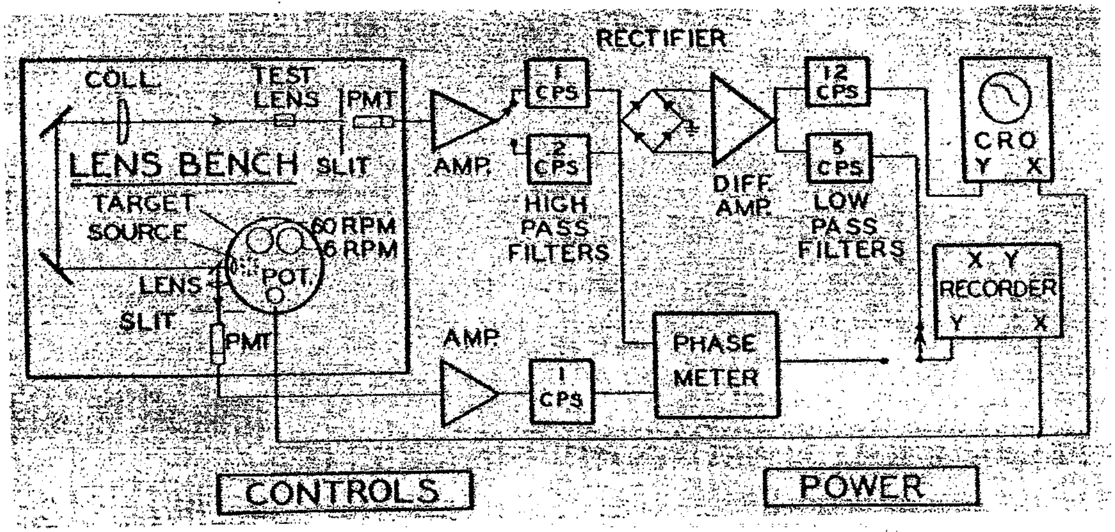


Figure 26.19 - Schematic diagram of the electronic system of the Herriott recording electronic lens bench. (From Jour. Optical Soc. America, W. Herriott 37; 472, 1947)

we suppose the optical system to accurately image a spatial frequency square wave such as a series of alternate dark and bright bars equally spaced, then the image may be represented as a spatial frequency series as,

$$F(x) = B_1 + \Delta B_1 \frac{4}{\pi} \left[\cos \left(2\pi n \frac{x}{\delta}\right) - \frac{1}{3} \cos 3\left(2\pi n \frac{x}{\delta}\right) + \frac{1}{5} \cos 5\left(2\pi n \frac{x}{\delta}\right) \ldots \right] \qquad (8)$$

Where x is the lateral coordinate and is defined as the width of a rectangle with the same area and height as the aperture flux distribution as shown in Figure 26.21 taken from Coltman (37). The square wave response factor is defined then as

$$r(n) = \frac{\Delta B_2/\Delta B_1}{B_2/B_1} \qquad (9)$$

and the sine wave response factor is defined as

$$R(n) = \frac{\Delta B_2/\Delta B_1}{B_2/B_1} \qquad (10)$$

26.4.1.2 It should be noted that there is a variation in the definition of response factors from author to author. This is clear if the reader will go back and check the definition of similar terms by Schade and D. Herriott. The end result in each case is essentially the same and one definition can be converted into another with no basic change in principle.

26.4.1.3 Coltman (38) shows that r(n) may be expressed in terms of the sine wave responses R(n) as given in equation (11).

$$r(n) = \frac{4}{\pi} \left[R(n) - \frac{R(3n)}{3} + \frac{R(5n)}{5} - \frac{R(7n)}{7} \ldots\ldots \right] \qquad (11)$$

solving for R(n) by successively subtracting series for $\frac{r(Kn)}{K}$ we can get,

$$R(n) = \frac{\pi}{4} \left[r(n) + \frac{r(3n)}{3} - \frac{r(5n)}{5} + \frac{r(7n)}{7} \ldots\ldots \right] \qquad (12)$$

The reader should see Coltman for the details. Suffice it to say that we have now expressed the sine wave response at a spatial frequency of n, the number of cycles in some unit distance. There are basically two ways of determining R(n). These will now be discussed.

26.4.2 The Coltman variable frequency square wave method.

26.4.2.1 The Coltman technique is similar in principle to the corresponding technique used in testing electrical amplifiers (39) with variable frequency square waves. Others such as Rosberry have also studied the method. It is usually found to be more trouble than it is worth to test electrical amplifiers this way, since if you have to vary the frequency of square wave, you might just as well vary the frequency of a sine wave and be done with it. In the optical case it is easier to vary the frequency of the square wave because spatial square waves can be made more easily than can spatial sine waves.

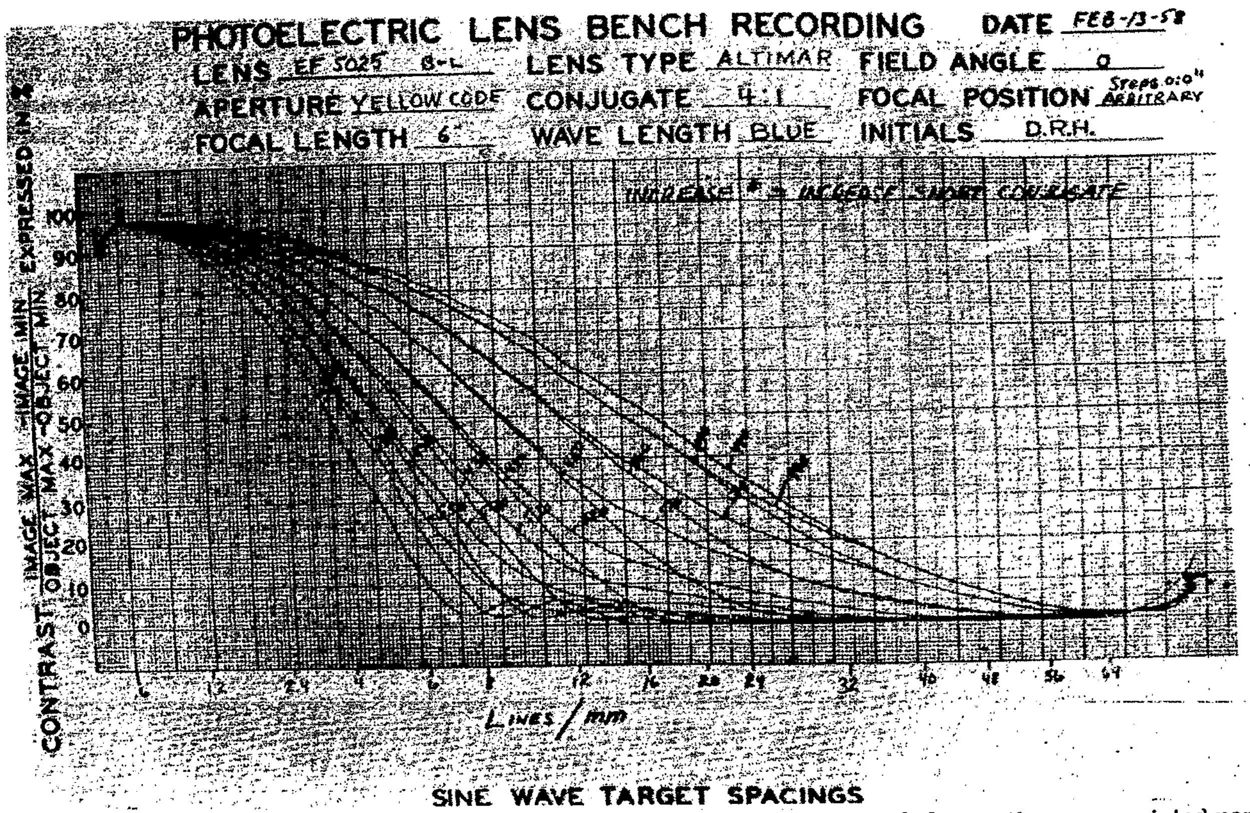
26.4.2.2 Coltman's method is similar, then, in principle to that discussed in Schade and Herriott's paper except for an analysis (40) that allows him to measure the sine wave response of the system by use of the more easily manufactured square waves. Not only is Coltman's article highly informative but it also gives an excellent discussion of the basis of the method and a specific example in the field of X-ray fluoroscopic work. Here the relative ease of studying systems in cascade by the sine wave method is shown and a discussion as to why sine wave targets are not used is given.

26.4.3 The fixed frequency square wave method.

26.4.3.1 Suppose that instead of variable spatial frequency square wave, we used a fixed spatial fequency square wave and get the higher frequency components by wave analysis of the electrical output of the photomultiplier tube. We now assume that the combination of scanning pinhole and associated photomultiplier circuit that transduces the spatial frequency to a temporal frequency spectrum in the image can be directly

---

(37) Coltman, JOSA 44, 468, (1954)
(38) ibid.
(39) Rosberry, A Correlation Investigation Between Photoelectric and Image Analysis, NBS Report No. 5799
(40) Loc. cit.,



PHOTOELECTRIC LENS BENCH RECORDING DATE FEB-13-58
LENS EF 5025 B-L LENS TYPE ALTIMAR FIELD ANGLE 0
APERTURE YELLOW CODE CONJUGATE 4:1 FOCAL POSITION Steps 0:0" ARBITRARY
FOCAL LENGTH 6" WAVE LENGTH BLUE INITIALS D.R.H.

Curves of contrast rendition measured through focus and recorded directly on preprinted paper.
Figure 26. 20 - Sample contrast rendition vs spatial frequency recording taken with the Herriott system.
(From Jour. Optical Soc. America, D. Herriott 48; 968, 1958)

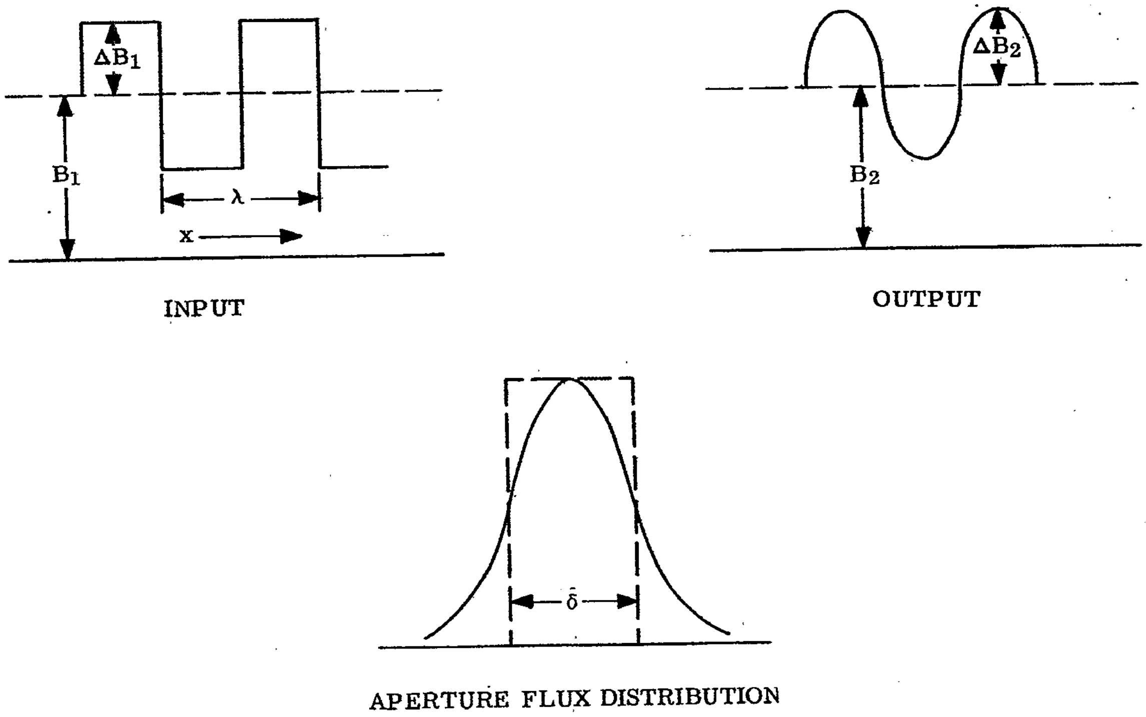


Figure 26. 21 - Quantities used in the Coltman definition of response factors.
(From Jour. Optical Soc. America, Coltman 44; 468, 1958)

related to the spatial frequency. Furthermore electrical tunable narrow band temporal frequency filters are standard items and have been for years. Therefore by feeding the output of the transducing element to a temporal frequency filter tuned to the fundamental of the square wave we can determine the sine wave response at that frequency. We then retune the filter to the next harmonic, record the output, etc. Account must be taken of the reduction in amplitude of the harmonics as given by the coefficients in equation (8).

26.4.3.2 It might occur to those versed in Fourier analysis that some other wave shape might be chosen such as a triangular wave that has all harmonics and not just the odd harmonics as in the case of the square wave. Such a wave could be used, but again the problem of production is such that it probably would be undesirable. Actually a single square wave target can suffice to cover almost any desired spatial frequency band - provided it is used with a minifying or magnifying system whose quality is far superior to that of the system being tested.

26.4.3.3 The difficulty involved with this method of square wave testing is that there is a phase shift associated with the tunable electrical filter. While there are ways to take this into account, they are rather complicated. Furthermore, it was assumed above that the percentage reduction of the amplitude as a function of frequency in the ideal image was known. This is true providing the detecting system is completely linear. For some systems, notably photographic ones, this may well not be so. Hence while we can get the sine wave response at any frequency, it may be difficult to relate it numerically to the response at other frequencies.

26.4.4 Automatic determination of power of an opthalmic lens by sine wave response.

26.4.4.1 In 1953 Gunter and Panetta (41) developed a method of applying the sine wave response criteria to the problem of automatically maintaining large aerial cameras in focus. The aerial camera aspects of the technique are of not so much interest to us here as is Gunter's definition of best focus used in connection with their analysis of the problem viz, "best focus is that point in image space where the spatial frequency response is an optimum within the bandwidth of information in which the observer is most interested." This definition is certainly a far cry from that usually found in optics and photography. It stems from the work of Schade rather than from that of the traditional treatments of Conrady etc.

26.4.4.2 Shortly thereafter Gunter (42), (43), (44) applied these same principles and this same definition of focus to the automatic determination of the power of opthalmic lenses. This was a research problem to see if the human factor could be removed in the routine inspection of opthalmic lenses. The women who customarily do this work are wont to get tired and their judgment varies. The first target was a square wave made by rotating a square cut gear as shown in Figure 26.22.

26.4.4.3 In the initial study the combination of SS and PM was moved along a lathe bed until the meter showed a maximum, the bandwidth of information having been selected by trial. Specifically this meant that a lens of say 2 diopters as judged by the eye was selected. This lens was placed in the test device and the temporal frequency filter adjusted until the meter output was maximum at a distance of exactly 50 cm. By checking with other standard lenses the variation of focal point as judged by the maximum meter response and the eye were shown to be well within commercial tolerances. A plot of meter response vs. focal position looked essentially the same as Figure 26.12.

26.4.4.4 The technique having been proved, the quest was now how to change the system so that a lens could be snapped into place and have the SS-PM unit automatically move so as to maximize the meter response i.e. move to the position of best focus. Important in the final system was a novel "square wave target" suggested by Hayes. The original square cut gear (seen from above) was modified as shown by the dotted lines in Figure 26.23 the hatched part of the gear remaining. The lens under test sees sequentially now edge a, b, a', b', a", b" etc. This is the equivalent of a square wave tipped at an angle of 45° in so far as L_X is concerned. The light from the source S was focussed midway between a and b so that the edges a and b appeared equally sharp to L_X. The action is simple. Referring to Figure 26.12 we see that near the peak the response curve as a function of focal length is quite symmetrical. An electronic circuit separated the responses from edges a and b and ordered a motor to adjust the position of the SS-PM unit until the responses were equal. The motor then stopped and the power of the lens was read directly from the scale. The system was easily more than sufficiently accurate. Modifications of the system were developed for specific purposes but the basic technique was unchanged.

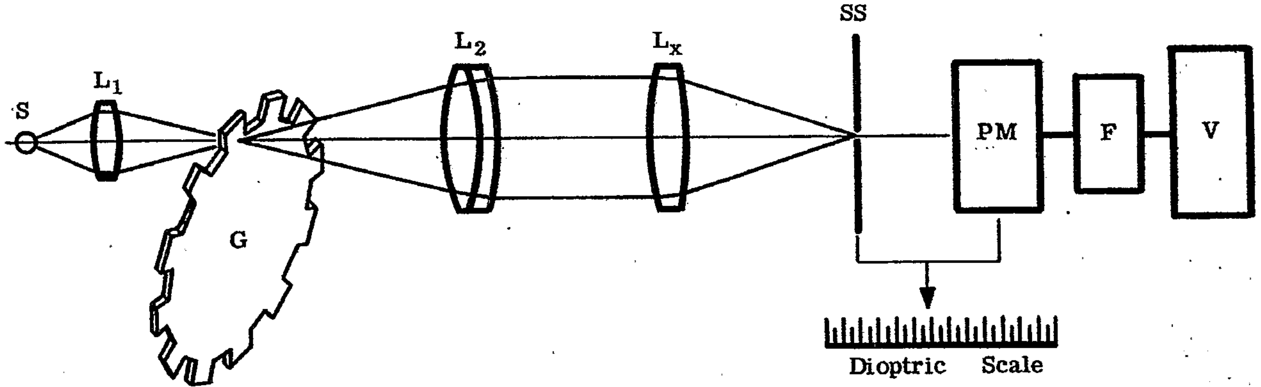
---

(41) Gunter and Panetta, An Automatic Electronic Focussing Device for Aerial Cameras, Boston University Optical Research Laboratory Technical Note 113, June, 1954

(42) Gunter, Whitney, Hayes. U. S. Patent 2897722, Electronic Lensometer.

(43) Gunter, Whitney, Hayes. U. S. Patent 2803995, Special Frequency Centering Device.

(44) Wing, Whitney, Hayes. U. S. Patent 2792748, Pyramid Centering Device.



S = source of light
L_1 = lens to focus light from S onto the teeth of G.
G = square cut gear rotated at 1800 rpm.
L_2 = a collimating lens.
L_x = the opthalmic lens under test.
SS = pinhole scanning aperture.
PM = photomultiplier tube and associated circuits
F = a tunable electric filter
V = voltmeter

Figure 26.22 - Basic square wave system for studying Opthalmic lens power.

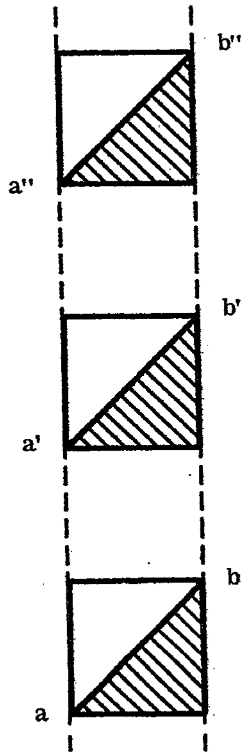


Figure 26.23 - The Hayes target for the American Optical automatic lens power measurement system.

Custodians:
Army - U.S. Army Munitions Command
Navy - Bureau of Ships
Air Force - Middletown Air Materiel Area

Preparing activity:
Army - U.S. Army Munitions Command

Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant
001	.001000	1.000	051	.051021	1.001	101	.101172	1.005	151	.151579	1.012
002	.002000	1.000	052	.052022	1.001	102	.102177	1.006	152	.152591	1.011
003	.003000	1.000	053	.053023	1.002	103	.103183	1.005	153	.153602	1.012
004	.004000	1.000	054	.054025	1.001	104	.104188	1.006	154	.154614	1.013
005	.005000	1.000	055	.055026	1.002	105	.105194	1.005	155	.155627	1.012
006	.006000	1.000	056	.056028	1.001	106	.106199	1.005	156	.156639	1.013
007	.007000	1.000	057	.057029	1.002	107	.107204	1.006	157	.157652	1.012
008	.008000	1.000	058	.058031	1.002	108	.108210	1.006	158	.158664	1.013
009	.009000	1.000	059	.059033	1.002	109	.109216	1.005	159	.159677	1.013
010	.010000	1.000	060	.060035	1.001	110	.110221	1.007	160	.160690	1.013
011	.011000	1.000	061	.061036	1.002	111	.111228	1.006	161	.161703	1.013
012	.012000	1.000	062	.062038	1.002	112	.112234	1.007	162	.162716	1.013
013	.013000	1.000	063	.063040	1.002	113	.113241	1.006	163	.163729	1.014
014	.014000	1.000	064	.064042	1.003	114	.114247	1.007	164	.164743	1.014
015	.015000	1.000	065	.065045	1.002	115	.115254	1.006	165	.165757	1.014
016	.016000	1.000	066	.066047	1.002	116	.116260	1.007	166	.166771	1.014
017	.017000	1.000	067	.067049	1.003	117	.117267	1.007	167	.167785	1.014
018	.018000	1.000	068	.068052	1.002	118	.118274	1.007	168	.168799	1.015
019	.019000	1.000	069	.069054	1.003	119	.119281	1.008	169	.169814	1.015
020	.020000	1.001	070	.070057	1.002	120	.120289	1.007	170	.170829	1.014
021	.021001	1.000	071	.071059	1.003	121	.121296	1.009	171	.171843	1.016
022	.022001	1.000	072	.072062	1.002	122	.122305	1.007	172	.172859	1.015
023	.023001	1.001	073	.073064	1.003	123	.123312	1.008	173	.173874	1.015
024	.024002	1.000	074	.074067	1.003	124	.124320	1.007	174	.174889	1.016
025	.025002	1.000	075	.075070	1.003	125	.125327	1.008	175	.175905	1.015
026	.026002	1.000	076	.076073	1.003	126	.126335	1.008	176	.176920	1.016
020	.027002	1.001	077	.077076	1.003	127	.127343	1.008	177	.177936	1.016
028	.028003	1.000	078	.078079	1.003	128	.128351	1.008	178	.178952	1.016
029	.029003	1.001	079	.079082	1.003	129	.129359	1.009	179	.179968	1.017
030	.030004	1.000	080	.080085	1.003	130	.130368	1.009	180	.180985	1.016
031	.031004	1.001	081	.081088	1.003	131	.131377	1.009	181	.182001	1.017
032	.032005	1.001	082	.082091	1.004	132	.132386	1.009	182	.183018	1.018
033	.033006	1.000	083	.083095	1.003	133	.133395	1.009	183	.184036	1.017
034	.034006	1.001	084	.084098	1.004	134	.134404	1.009	184	.185053	1.017
035	.035007	1.000	085	.085102	1.004	135	.135413	1.009	185	.186070	1.018
036	.036007	1.001	086	.086106	1.003	136	.136422	1.010	186	.187088	1.018
037	.037008	1.001	087	.087109	1.004	137	.137432	1.010	187	.188106	1.018
038	.038009	1.000	088	.088113	1.004	138	.138442	1.009	188	.189124	1.018
039	.039009	1.001	089	.089117	1.004	139	.139451	1.010	189	.190142	1.019
040	.040010	1.001	090	.090121	1.004	140	.140461	1.010	190	.191161	1.019
041	.041011	1.001	091	.091125	1.004	141	.141471	1.010	191	.192180	1.019
042	.042012	1.001	092	.092129	1.005	142	.142481	1.010	192	.193199	1.019
043	.043013	1.001	093	.093134	1.004	143	.143491	1.011	193	.194218	1.019
044	.044014	1.001	094	.094138	1.004	144	.144502	1.010	194	.195237	1.020
045	.045015	1.001	095	.095142	1.005	145	.145512	1.011	195	.196257	1.019
046	.046016	1.001	096	.096147	1.005	146	.146523	1.011	196	.197276	1.020
047	.047017	1.001	097	.097152	1.005	147	.147534	1.011	197	.198296	1.020
048	.048018	1.001	098	.098157	1.004	148	.148545	1.011	198	.199316	1.021
049	.049019	1.001	099	.099161	1.006	149	.149556	1.012	199	.200337	1.020
050	.050020	1.001	100	.100167	1.005	150	.150568	1.011	200	.201357	1.021

Table 1 - Sine - Angle Conversion Table (Sheet 1 of 5)

Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant
201	.202378	1.021	251	.253717	1.033	301	.305741	1.049	351	.358638	1.068
202	.203399	1.021	252	.254746	1.033	302	.306790	1.049	352	.359706	1.068
203	.204420	1.021	253	.255779	1.035	303	.307839	1.049	353	.360774	1.069
204	.205441	1.022	254	.256814	1.034	304	.308888	1.050	354	.361843	1.070
205	.206463	1.022	255	.257848	1.034	305	.309938	1.050	355	.362913	1.070
206	.207485	1.022	256	.258882	1.035	306	.310988	1.051	356	.363983	1.070
207	.208507	1.022	257	.259917	1.034	307	.312039	1.051	357	.365053	1.071
208	.209529	1.022	258	.260951	1.035	308	.313090	1.051	358	.366124	1.072
209	.210551	1.023	259	.261986	1.035	309	.314141	1.052	359	.367196	1.072
210	.211574	1.023	260	.263021	1.036	310	.315193	1.052	360	.368268	1.071
211	.212597	1.023	261	.264057	1.036	311	.316245	1.052	361	.369339	1.073
212	.213620	1.023	262	.265093	1.037	312	.317297	1.053	362	.370412	1.073
213	.214643	1.024	263	.266130	1.036	313	.318350	1.053	363	.371485	1.074
214	.215667	1.024	264	.267166	1.037	314	.319403	1.053	364	.372559	1.074
215	.216691	1.024	265	.268203	1.038	315	.320456	1.054	365	.373633	1.074
216	.217715	1.024	266	.269241	1.037	316	.321510	1.054	366	.374707	1.075
217	.218739	1.025	267	.270278	1.038	317	.322564	1.054	367	.375782	1.075
218	.219764	1.024	268	.271316	1.038	318	.323618	1.055	368	.376857	1.076
219	.220788	1.025	269	.272354	1.039	319	.324673	1.056	369	.377933	1.076
220	.221813	1.026	270	.273393	1.038	320	.325729	1.056	370	.379009	1.076
221	.222839	1.025	271	.274431	1.039	321	.326785	1.056	371	.380085	1.077
222	.223864	1.026	272	.275470	1.039	322	.327841	1.056	372	.381162	1.078
223	.224890	1.026	273	.276509	1.040	323	.328897	1.057	373	.382240	1.078
224	.225916	1.026	274	.277549	1.040	324	.329954	1.058	374	.383318	1.078
225	.226942	1.027	275	.278589	1.040	325	.331012	1.057	375	.384396	1.079
226	.227969	1.026	276	.279629	1.041	326	.332069	1.058	376	.385475	1.079
227	.228995	1.027	277	.280670	1.041	327	.333127	1.058	377	.386554	1.081
228	.230022	1.027	278	.281711	1.041	328	.334185	1.059	378	.387635	1.080
229	.231049	1.028	279	.282752	1.041	329	.335244	1.059	379	.388715	1.081
230	.232077	1.027	280	.283793	1.042	330	.336303	1.060	380	.389796	1.082
231	.233104	1.028	281	.284835	1.042	331	.337363	1.060	381	.390878	1.081
232	.234132	1.029	282	.285877	1.043	332	.338423	1.060	382	.391959	1.082
233	.235161	1.028	283	.286920	1.043	333	.339483	1.061	383	.383041	1.083
234	.236189	1.029	284	.287963	1.043	334	.340544	1.061	384	.394124	1.083
235	.237218	1.028	285	.289006	1.043	335	.341605	1.061	385	.395207	1.084
236	.238246	1.030	286	.290049	1.044	336	.342666	1.062	386	.396291	1.084
237	.239276	1.029	287	.291093	1.044	337	.343728	1.063	387	.397375	1.085
238	.240305	1.030	288	.292137	1.044	338	.344791	1.062	388	.398460	1.085
239	.241335	1.030	289	.293181	1.045	339	.345853	1.063	389	.399545	1.086
240	.242365	1.030	290	.294226	1.045	340	.346916	1.064	390	.400631	1.086
241	.243395	1.031	291	.295271	1.045	341	.347980	1.064	391	.401717	1.087
242	.244426	1.031	292	.296316	1.047	342	.349044	1.064	392	.402804	1.087
243	.245457	1.031	293	.297363	1.046	343	.350108	1.065	393	.403891	1.088
244	.246488	1.031	294	.298409	1.046	344	.351173	1.065	394	.404979	1.088
245	.247519	1.032	295	.299455	1.047	345	.352238	1.065	395	.406067	1.088
246	.248551	1.032	296	.300502	1.047	346	.353303	1.067	396	.407155	1.090
247	.249583	1.032	297	.301549	1.048	348	.354370	1.067	397	.408245	1.090
248	.250615	1.032	298	.302597	1.048	348	.355437	1.066	398	.409335	1.090
249	.251647	1.033	299	.303645	1.047	349	.356503	1.067	399	.410425	1.091
250	.252680	1.033	300	.304692	1.049	350	.357570	1.068	400	.411516	1.091

Table 1 - Sine - Angle Conversion Table (Sheet 2 of 5)

Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant
401	.412607	1.092	451	.467885	1.121	501	.524754	1.156	551	.583562	1.199
402	.413699	1.092	452	.469006	1.121	502	.525910	1.156	552	.584761	1.200
403	.414791	1.093	453	.470127	1.122	503	.527066	1.158	553	.585961	1.200
404	.415884	1.094	454	.471249	1.123	504	.528224	1.158	554	.587161	1.202
405	.416978	1.094	455	.472372	1.123	505	.529382	1.159	555	.588363	1.203
406	.418072	1.095	456	.473495	1.124	506	.530541	1.160	556	.589566	1.203
407	.419167	1.095	457	.474619	1.125	507	.531701	1.160	557	.590769	1.205
408	.420262	1.096	458	.475744	1.125	508	.532861	1.162	558	.591974	1.205
409	.421358	1.096	459	.476869	1.126	509	.534023	1.162	559	.593179	1.207
410	.422454	1.096	460	.477995	1.127	510	.535185	1.163	560	.594386	1.207
411	.423550	1.097	461	.479122	1.127	511	.536348	1.163	561	.595593	1.209
412	.424647	1.098	462	.480249	1.128	512	.537511	1.165	562	.596802	1.209
413	.425745	1.099	463	.481377	1.128	513	.538676	1.165	563	.598011	1.211
414	.426844	1.099	464	.482505	1.130	514	.539841	1.167	564	.599222	1.211
415	.427943	1.100	465	.483635	1.130	515	.541008	1.167	565	.600433	1.213
416	.429043	1.099	466	.484765	1.131	516	.542175	1.167	566	.601646	1.213
417	.430142	1.101	467	.485895	1.132	517	.543342	1.169	567	.602859	1.215
418	.431243	1.101	468	.487026	1.133	518	.544511	1.170	568	.604074	1.215
419	.432344	1.101	469	.488158	1.133	519	.545681	1.170	569	.605289	1.217
420	.433445	1.103	470	.489291	1.133	520	.546851	1.171	570	.606506	1.217
421	.434548	1.102	471	.490424	1.134	521	.548022	1.172	571	.607723	1.219
422	.435650	1.103	472	.491558	1.135	522	.549194	1.173	572	.608942	1.220
423	.436753	1.104	473	.492693	1.135	523	.550367	1.173	573	.610162	1.220
424	.437857	1.104	474	.493828	1.136	524	.551540	1.175	574	.611382	1.222
425	.438961	1.105	475	.494964	1.137	525	.552715	1.175	575	.612604	1.223
426	.440066	1.106	476	.496101	1.137	526	.553890	1.177	576	.613827	1.224
427	.441172	1.106	477	.497238	1.138	527	.555067	1.177	577	.615051	1.225
428	.442278	1.107	478	.498376	1.139	528	.556244	1.178	578	.616276	1.226
429	.443385	1.107	479	.499515	1.140	529	.557422	1.179	579	.617502	1.227
430	.444492	1.108	480	.500655	1.140	530	.558601	1.179	580	.618729	1.228
431	.445600	1.109	481	.501795	1.141	531	.559780	1.181	581	.619957	1.229
432	.446709	1.108	482	.502936	1.142	532	.560961	1.181	582	.621186	1.230
433	.447817	1.110	483	.504078	1.142	533	.562142	1.182	583	.622416	1.232
434	.448927	1.111	484	.505220	1.143	534	.563324	1.184	584	.623648	1.232
435	.450038	1.110	485	.506363	1.144	535	.564508	1.184	585	.624880	1.234
436	.451148	1.112	486	.507507	1.145	536	.565692	1.185	586	.626114	1.234
437	.452260	1.112	487	.508652	1.145	537	.566877	1.186	587	.627348	1.236
438	.453372	1.113	488	.509797	1.146	538	.568063	1.186	588	.628584	1.237
439	.454485	1.113	489	.510943	1.147	539	.569249	1.188	589	.629821	1.238
440	.455598	1.113	490	.512090	1.147	540	.570437	1.189	590	.631059	1.239
441	.456711	1.115	491	.513237	1.148	541	.571626	1.189	591	.632298	1.240
442	.457826	1.115	492	.514385	1.149	542	.572815	1.191	592	.633538	1.242
443	.458941	1.116	493	.515534	1.150	543	.574006	1.191	593	.634780	1.242
444	.460057	1.116	494	.516684	1.151	544	.575197	1.192	594	.636022	1.244
445	.461173	1.117	495	.517835	1.151	545	.576389	1.193	595	.637266	1.245
446	.462290	1.113	496	.518986	1.152	546	.577582	1.194	596	.638511	1.245
447	.463408	1.119	497	.520138	1.153	547	.578776	1.195	597	.639756	1.247
448	.464527	1.118	498	.521291	1.153	548	.579971	1.196	598	.641003	1.249
449	.465645	1.120	499	.522444	1.155	549	.581167	1.197	599	.642252	1.249
450	.466765	1.120	500	.523599	1.155	550	.582364	1.198	600	.643501	1.251

Table 1 - Sine - Angle Conversion Table (Sheet 3 of 5)

Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant
601	.644752	1.251	651	.708901	1.318	701	.776799	1.403	751	.849575	1.516
602	.646003	1.253	652	.710219	1.320	702	.778202	1.405	752	.851091	1.518
603	.647256	1.255	653	.711539	1.321	703	.779607	1.407	753	.852609	1.521
604	.648511	1.256	654	.712860	1.323	704	.781014	1.409	754	.854130	1.524
605	.649766	1.256	655	.714183	1.324	705	.782423	1.411	755	.855654	1.526
606	.651022	1.258	656	.715507	1.325	706	.783334	1.413	756	.857180	1.529
607	.652280	1.259	657	.716832	1.328	707	.785247	1.415	757	.858709	1.532
608	.653539	1.260	658	.718160	1.329	708	.786662	1.417	758	.860241	1.535
609	.654799	1.262	659	.719489	1.330	709	.788079	1.419	759	.861776	1.537
610	.656061	1.262	660	.720819	1.332	710	.789498	1.421	760	.863313	1.540
611	.657323	1.264	661	.722151	1.333	711	.790919	1.423	761	.864853	1.543
612	.658587	1.265	662	.723484	1.335	712	.792342	1.426	762	.866396	1.546
613	.659852	1.266	663	.724819	1.337	713	.793768	1.427	763	.867942	1.548
614	.661118	1.268	664	.726156	1.338	714	.795195	1.429	764	.869490	1.551
615	.662386	1.269	665	.727494	1.340	715	.796624	1.432	765	.871041	1.555
616	.663655	1.270	666	.728834	1.341	716	.798056	1.433	766	.872596	1.557
617	.664925	1.271	667	.730175	1.343	717	.799489	1.436	767	.874153	1.559
618	.666196	1.273	668	.731518	1.345	718	.800925	1.438	768	.875712	1.563
619	.667469	1.274	669	.732863	1.346	719	.802363	1.439	769	.877275	1.566
620	.668743	1.275	670	.734209	1.348	720	.803802	1.442	770	.878841	1.569
621	.670018	1.276	671	.735557	1.349	721	.805244	1.445	771	.880410	1.572
622	.671294	1.278	672	.736906	1.351	722	.806689	1.446	772	.881982	1.574
623	.672572	1.279	673	.738257	1.353	723	.808135	1.449	773	.883556	1.578
624	.673851	1.281	674	.739610	1.355	724	.809584	1.450	774	.885134	1.581
625	.675132	1.281	675	.740965	1.356	725	.811034	1.453	775	.886715	1.584
626	.676413	1.283	676	.742321	1.358	726	.812487	1.456	776	.888299	1.587
627	.677696	1.284	677	.743679	1.359	727	.813943	1.457	777	.889886	1.590
628	.678980	1.286	678	.745038	1.362	728	.815400	1.460	778	.891476	1.593
629	.680266	1.287	679	.746400	1.363	729	.816860	1.462	779	.893069	1.597
630	.681553	1.289	680	.747763	1.364	730	.818322	1.464	780	.894666	1.599
631	.682842	1.289	681	.749127	1.367	731	.819786	1.467	781	.896265	1.603
632	.684131	1.291	682	.750494	1.368	732	.821253	1.469	782	.897868	1.606
633	.685422	1.293	683	.751862	1.370	733	.822722	1.471	783	.899474	1.610
634	.686715	1.293	684	.753232	1.372	734	.824693	1.474	784	.901084	1.612
635	.688008	1.296	685	.754604	1.373	735	.825667	1.476	785	.902696	1.616
636	.689304	1.296	686	.755977	1.376	736	.827143	1.478	786	.904312	1.619
637	.690600	1.298	687	.757353	1.377	737	.828621	1.481	787	.905931	1.623
638	.691898	1.299	688	.758730	1.378	738	.830102	1.483	788	.907554	1.626
639	.693197	1.301	689	.760108	1.381	739	.831585	1.485	789	.909180	1.629
640	.694498	1.302	690	.761489	1.383	740	.833070	1.488	790	.910809	1.633
641	.695800	1.304	691	.762872	1.384	741	.834558	1.491	791	.912442	1.636
642	.697104	1.305	692	.764256	1.386	742	.836049	1.493	792	.914078	1.640
643	.698409	1.306	693	.765642	1.388	743	.837542	1.495	793	.915718	1.643
644	.699715	1.308	694	.767030	1.390	744	.839037	1.498	794	.917361	1.647
645	.701023	1.310	695	.768420	1.392	745	.840535	1.500	795	.919008	1.650
646	.702333	1.310	696	.769812	1.393	746	.842035	1.503	796	.920658	1.654
647	.703643	1.313	697	.771205	1.396	747	.843538	1.505	797	.922312	1.657
648	.704956	1.313	698	.772601	1.397	748	.845043	1.509	798	.923969	1.661
649	.706269	1.315	699	.773998	1.400	749	.846552	1.510	799	.925630	1.665
650	.707584	1.317	700	.775398	1.401	750	.848062	1.513	800	.927295	1.669

Table 1 - Sine - Angle Conversion Table (Sheet,4 of 5)

Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant	Sine	Angle	Interp. Constant
801	.928964	1.672	851	1.017887	1.907	901	1.122069	2.311	951	1.256454	3.251
802	.930636	1.676	852	1.019794	1.913	902	1.124380	2.322	952	1.259705	3.284
803	.932312	1.680	853	1.021707	1.919	903	1.126702	2.333	953	1.262989	3.317
804	.933992	1.683	854	1.023626	1.925	904	1.129035	2.345	954	1.266306	3.354
805	.935675	1.688	855	1.025551	1.931	905	1.131380	2.356	955	1.269660	3.390
806	.937363	1.691	856	1.027482	1.938	906	1.133736	2.368	956	1.273050	3.428
807	.939054	1.696	857	1.029420	1.943	907	1.136104	2.381	957	1.276478	3.467
808	.940750	1.699	858	1.031363	1.950	908	1.138485	2.393	958	1.279945	3.508
809	.942449	1.703	859	1.033313	1.957	909	1.140878	2.406	959	1.283453	3.550
810	.944152	1.707	860	1.035270	1.963	910	1.143282	2.419	960	1.287003	3.593
811	.945859	1.712	861	1.037233	1.969	911	1.145703	2.431	961	1.290596	3.639
812	.947571	1.715	862	1.039202	1.976	912	1.148134	2.444	962	1.294235	3.686
813	.949286	1.720	863	1.041178	1.983	913	1.150578	2.458	963	1.297921	3.736
814	.951006	1.723	864	1.043161	1.989	914	1.153036	2.472	964	1.301657	3.787
815	.952729	1.728	865	1.045150	1.997	915	1.155508	2.486	965	1.305444	3.840
816	.954457	1.732	866	1.047147	2.003	916	1.157994	2.500	966	1.309284	3.896
817	.956189	1.736	867	1.049150	2.010	917	1.160494	2.514	967	1.313180	3.955
818	.957925	1.741	868	1.051160	2.018	918	1.163008	2.529	968	1.317135	4.016
819	.959666	1.745	869	1.053178	2.024	919	1.165537	2.544	969	1.321151	4.080
820	.961411	1.749	870	1.055202	2.032	920	1.168081	2.559	970	1.325231	4.148
821	.963160	1.754	871	1.057234	2.039	921	1.168081	2.575	971	1.329379	4.219
822	.964914	1.758	872	1.059273	2.047	922	1.173215	2.591	972	1.333598	4.294
823	.966672	1.763	873	1.061320	2.054	923	1.175806	2.607	973	1.337892	4.373
824	.968435	1.767	874	1.063374	2.062	924	1.178413	2.623	974	1.342265	4.457
825	.970202	1.772	875	1.065436	2.069	925	1.181036	2.640	975	1.346722	4.545
826	.971974	1.776	876	1.067505	2.078	926	1.183676	2.658	976	1.351267	4.640
827	.973750	1.782	877	1.069583	2.085	927	1.186334	2.675	977	1.355907	4.742
828	.975532	1.785	878	1.071668	2.093	928	1.189009	2.693	978	1.360649	4.849
829	.977317	1.791	879	1.073761	2.101	929	1.191702	2.711	979	1.365498	4.964
830	.979108	1.795	880	1.075862	2.110	930	1.194413	2.730	980	1.370462	5.089
831	.980903	1.800	881	1.077972	2.117	931	1.197143	2.749	981	1.375551	5.223
832	.982703	1.805	882	1.080089	2.127	932	1.199892	2.769	982	1.380774	5.370
833	.984508	1.810	883	1.082216	2.135	933	1.202661	2.789	983	1.386144	5.529
834	.986318	1.815	884	1.084351	2.143	934	1.205450	2.809	984	1.391673	5.702
835	.988133	1.820	885	1.086494	2.152	935	1.208259	2.831	985	1.397375	5.894
836	.989953	1.825	886	1.088646	2.161	936	1.211090	2.852	986	1.403269	6.108
837	.991778	1.830	887	1.090807	2.170	937	1.213942	2.874	987	1.409377	6.345
838	.993608	1.835	888	1.092977	2.180	938	1.216816	2.896	988	1.415722	6.614
839	.995443	1.840	889	1.095157	2.188	939	1.219712	2.919	989	1.422336	6.921
840	.997283	1.846	890	1.097345	2.198	940	1.222631	2.943	990	1.429257	7.275
841	.999129	1.851	891	1.099543	2.207	941	1.225574	2.967	991	1.436532	7.689
842	1.000980	1.856	892	1.101750	2.218	942	1.228541	2.992	992	1.444221	8.186
843	1.002862	1.862	893	1.103968	2.226	943	1.231533	3.018	993	1.452407	8.791
844	1.004698	1.867	894	1.106194	2.238	944	1.234551	3.044	994	1.461198	9.557
845	1.006565	1.873	895	1.108432	2.246	945	1.237595	3.071	995	1.470755	10.569
846	1.008438	1.878	896	1.110678	2.257	946	1.240666	3.099	996	1.481324	11.993
847	1.010316	1.884	897	1.112935	2.268	947	1.243765	3.127	997	1.493317	14.225
848	1.012200	1.890	898	1.115203	2.278	948	1.246892	3.157	998	1.507542	18.529
849	1.014090	1.895	899	1.117481	2.289	949	1.250049	3.187	999	1.526071	44.725
850	1.015985	1.902	900	1.119770	2.299	950	1.253236	3.218	1000	1.570796	

Table 1 - Sine - Angle Conversion Table (Sheet 5 of 5)

Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant
001	.001000	1.000	051	.050978	.999	101	.100828	.995	151	.150427	.988
002	.002000	1.000	052	.051977	.998	102	.101823	.995	152	.151415	.989
003	.003000	1.000	053	.052975	.999	103	.102818	.995	153	.152404	.988
004	.004000	1.000	054	.053974	.998	104	.103813	.994	154	.153392	.988
005	.005000	1.000	055	.054972	.999	105	.104807	.995	155	.154380	.988
006	.006000	1.000	056	.055971	.998	106	.105802	.994	156	.155368	.988
007	.007000	1.000	057	.056969	.998	107	.106796	.994	157	.156356	.987
008	.008000	1.000	058	.057967	.999	108	.107790	.994	158	.157343	.988
009	.009000	1.000	059	.058966	.998	109	.108784	.994	159	.158331	.987
010	.010000	1.000	060	.059964	.998	110	.109778	.994	160	.159318	.987
011	.011000	1.000	061	.060962	.998	111	.110772	.994	161	.160305	.987
012	.012000	1.000	062	.061960	.998	112	.111766	.994	162	.161292	.987
013	.013000	1.000	063	.062958	.998	113	.112760	.993	163	.162279	.987
014	.014000	.999	064	.063956	.998	114	.113753	.994	164	.163266	.986
015	.014999	1.000	065	.064954	.998	115	.114747	.993	165	.164252	.987
016	.015999	1.000	066	.065952	.998	116	.115740	.993	166	.165239	.986
017	.016999	1.000	067	.066950	.998	117	.116733	.993	167	.166225	.986
018	.017999	1.000	068	.067948	.997	118	.117726	.993	168	.167211	.986
019	.018999	1.000	069	.068945	.998	119	.118719	.993	169	.168197	.985
020	.019999	.999	070	.069943	.997	120	.119712	.993	170	.169182	.985
021	.020998	1.000	071	.070940	.998	121	.120705	.993	171	.170167	.986
022	.021998	1.000	072	.071938	.997	122	.121698	.992	172	.171153	.985
023	.022998	1.000	073	.072935	.997	123	.122690	.992	173	.172138	.985
024	.023998	.999	074	.073932	.998	124	.123682	.993	174	.173123	.985
025	.024997	1.000	075	.074930	.997	125	.124675	.992	175	.174108	.985
026	.025997	1.000	076	.075927	.997	126	.125667	.992	176	.175093	.984
027	.026997	.999	077	.076924	.997	127	.126659	.992	177	.176077	.985
028	.027996	1.000	078	.077921	.997	128	.127651	.992	178	.177062	.984
029	.028996	1.000	079	.078918	.997	129	.128643	.991	179	.178046	.984
030	.029996	.999	080	.079915	.996	130	.129634	.992	180	.179030	.983
031	.030995	1.000	081	.080911	.997	131	.130626	.991	181	.180013	.984
032	.031995	.999	082	.081908	.997	132	.131617	.991	182	.180997	.983
033	.032994	.999	083	.082905	.996	133	.132608	.991	183	.181980	.984
034	.033993	1.000	084	.083901	.997	134	.133599	.991	184	.182964	.982
035	.034993	.999	085	.084898	.996	135	.134590	.991	185	.183946	.983
036	.035992	1.000	086	.085894	.996	136	.135581	.991	186	.184929	.983
037	.036992	.999	087	.086890	.996	137	.136572	.990	187	.185912	.983
038	.037991	.999	088	.087886	.997	138	.137562	.991	188	.186895	.982
039	.038990	.999	089	.088883	.996	139	.138553	.990	189	.187877	.982
040	.039989	1.000	090	.089879	.995	140	.139543	.990	190	.188859	.982
041	.040989	.999	091	.090874	.996	141	.140533	.990	191	.189841	.982
042	.041988	.999	092	.091870	.996	142	.141523	.990	192	.190823	.981
043	.042987	.999	093	.092866	.996	143	.142513	.990	193	.191804	.981
044	.043986	.999	094	.093862	.995	144	.143503	.989	194	.192785	.982
045	.044985	.999	095	.094857	.996	145	.144492	.990	195	.193767	.980
046	.045984	.999	096	.095853	.995	146	.145482	.989	196	.194747	.981
047	.046983	.999	097	.096848	.995	147	.146471	.989	197	.195728	.981
048	.047982	.998	098	.097843	.995	148	.147460	.989	198	.196709	.980
049	.048980	.999	099	.098838	.995	149	.148449	.989	199	.197689	.980
050	.049979	.999	100	.099833	.995	150	.149438	.989	200	.198669	.980

Table 2 - Angle - Sine Conversion Table (Sheet 1 of 5)

Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant
201	.199649	.980	251	.248373	.968	301	.296475	.955	351	.343837	.939
202	.200629	.979	252	.249341	.969	302	.297430	.955	352	.344776	.938
203	.201608	.980	253	.250310	.968	303	.298385	.954	353	.345714	.939
204	.202588	.979	254	.251278	.967	304	.299339	.954	354	.346653	.937
205	.203567	.979	255	.252245	.968	305	.300293	.954	355	.347590	.938
206	.204546	.979	256	.253213	.967	306	.301247	.953	356	.348528	.937
207	.205525	.978	257	.254180	.967	307	.302200	.953	357	.349465	.937
208	.206503	.979	258	.255147	.967	308	.303153	.953	358	.350402	.936
209	.207482	.978	259	.256114	.967	309	.304106	.953	359	.351338	.936
210	.208460	.978	260	.257081	.966	310	.305059	.952	360	.352274	.936
211	.209438	.978	261	.258047	.966	311	.306011	.952	361	.353210	.935
212	.210416	.977	262	.259013	.966	312	.306963	.951	362	.354145	.935
213	.211393	.977	263	.259979	.965	313	.307914	.952	363	.355080	.935
214	.212370	.977	264	.260944	.965	314	.308866	.950	364	.356015	.934
215	.213347	.977	265	.261909	.965	315	.309816	.951	365	.356949	.934
216	.214324	.977	266	.262874	.965	316	.310767	.950	366	.357883	.934
217	.215301	.976	267	.263839	.964	317	.311717	.950	367	.358817	.933
218	.216277	.977	268	.264803	.965	318	.312667	.950	368	.359750	.933
219	.217254	.976	269	.265768	.963	319	.313617	.950	369	.360683	.932
220	.218230	.975	270	.266731	.964	320	.314567	.949	370	.361615	.933
221	.219205	.976	271	.267695	.963	321	.315516	.948	371	.362548	.931
222	.220181	.975	272	.268658	.964	322	.316464	.949	372	.363479	.932
223	.221156	.975	273	.269622	.962	323	.317413	.948	373	.364411	.931
224	.222131	.975	274	.270584	.963	324	.318361	.948	374	.365342	.931
225	.223106	.975	275	.271547	.962	325	.319309	.947	375	.366273	.930
226	.224081	.975	276	.272509	.962	326	.320256	.947	376	.367203	.930
227	.225056	.974	277	.273471	.962	327	.321203	.947	377	.368133	.929
228	.226030	.974	278	.274433	.961	328	.322150	.947	378	.369062	.930
229	.227004	.974	279	.275394	.962	329	.323097	.946	379	.369992	.928
230	.227978	.973	280	.276356	.961	330	.324043	.946	380	.370920	.929
231	.228951	.973	281	.277317	.960	331	.324989	.945	381	.371849	.928
232	.229924	.973	282	.278277	.961	332	.325934	.946	382	.372777	.928
233	.230897	.973	283	.279238	.960	333	.326880	.945	383	.373705	.927
234	.231870	.973	284	.280198	.959	334	.327825	.944	384	.374632	.927
235	.232843	.972	285	.281157	.960	335	.328769	.944	385	.375559	.927
236	.233815	.973	286	.282117	.959	336	.329713	.944	386	.376486	.926
237	.234788	.971	287	.283076	.959	337	.330657	.944	387	.377412	.926
238	.235759	.972	288	.284035	.959	338	.331601	.943	388	.378338	.925
239	.236731	.972	289	.284994	.958	339	.332544	.943	389	.379263	.925
240	.237703	.971	290	.285952	.958	340	.333487	.943	390	.380188	.925
241	.238674	.971	291	.286910	.958	341	.334430	.942	391	.381113	.924
242	.239645	.971	292	.287868	.958	342	.335372	.942	392	.382037	.924
243	.240616	.970	293	.288826	.957	343	.336314	.941	393	.382961	.924
244	.241586	.970	294	.289783	.957	344	.337255	.942	394	.383885	.923
245	.242556	.970	295	.290740	.957	345	.338197	.941	395	.384808	.923
246	.243526	.970	296	.291697	.956	346	.339138	.940	396	.385731	.922
247	.244496	.970	297	.292653	.956	347	.340078	.940	397	.386653	.922
248	.245466	.969	298	.293609	.956	348	.341018	.940	398	.387575	.922
249	.246435	.969	299	.294565	.955	349	.341958	.940	399	.388497	.921
250	.247404	.969	300	.295520	.955	350	.342898	.939	400	.389418	.921

Table 2 - Angle - Sine Conversion Table (Sheet 2 of 5)

Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant
401	.390339	.921	451	.435866	.900	501	.480303	.877	551	.523539	.852
402	.391260	.920	452	.436766	.899	502	.481180	.876	552	.524391	.851
403	.392180	.919	453	.437665	.899	503	.482056	.876	553	.525242	.851
404	.393099	.920	454	.438564	.898	504	.482932	.875	554	.526093	.850
405	.394019	.919	455	.439462	.898	505	.483807	.875	555	.526943	.850
406	.394938	.918	456	.440360	.898	506	.484682	.875	556	.527793	.849
407	.395856	.918	457	.441258	.897	507	.485557	.874	557	.528642	.849
408	.396774	.918	458	.442155	.897	508	.486431	.873	558	.529491	.848
409	.397692	.917	459	.443052	.896	509	.487304	.873	559	.530339	.847
410	.398609	.917	460	.443948	.896	510	.488177	.873	560	.531186	.847
411	.399526	.917	461	.444844	.895	511	.489050	.872	561	.532033	.847
412	.400443	.916	462	.445739	.895	512	.489922	.871	562	.532880	.846
413	.401359	.916	463	.446634	.895	513	.490793	.871	563	.533726	.845
414	.402275	.915	464	.447529	.894	514	.491664	.871	564	.534571	.845
415	.403190	.915	465	.448423	.893	515	.492535	.870	565	.535416	.844
416	.404105	.914	466	.449316	.894	516	.493405	.869	566	.536260	.844
417	.405019	.914	467	.450210	.892	517	.494274	.870	567	.537104	.843
418	.405933	.914	468	.451102	.892	518	.495144	.868	568	.537947	.843
419	.406847	.913	469	.451994	.892	519	.496012	.868	569	.538790	.842
420	.407760	.913	470	.452886	.892	520	.496880	.868	570	.539632	.842
421	.408673	.913	471	.453778	.891	521	.497748	.867	571	.540474	.841
422	.409586	.912	472	.454669	.890	522	.498615	.866	572	.541315	.840
423	.410498	.912	473	.455559	.890	523	.499481	.866	573	.542155	.840
424	.411410	.911	474	.456449	.889	524	.500347	.866	574	.542995	.840
425	.412321	.911	475	.457338	.890	525	.501213	.865	575	.543835	.839
426	.413232	.910	476	.458228	.888	526	.502078	.865	576	.544674	.838
427	.414142	.910	477	.459116	.888	527	.502943	.864	577	.545512	.838
428	.415052	.910	478	.460004	.888	528	.503807	.863	578	.546350	.837
429	.415962	.909	479	.460892	.887	529	.504670	.863	579	.547187	.837
430	.416871	.909	480	.461779	.887	530	.505533	.863	580	.548024	.836
431	.417780	.908	481	.462666	.886	531	.506396	.862	581	.548860	.836
432	.418688	.908	482	.463552	.886	532	.507258	.861	582	.549696	.835
433	.419596	.907	483	.464438	.885	533	.508119	.862	583	.550531	.834
434	.420503	.907	484	.465323	.885	534	.508981	.860	584	.551365	.834
435	.421410	.907	485	.466208	.885	535	.509841	.860	585	.552199	.834
436	.422317	.906	486	.467093	.884	536	.510701	.860	586	.553033	.833
437	.423223	.906	487	.467977	.883	537	.511561	.859	587	.553866	.832
438	.424129	.906	488	.468860	.883	538	.512420	.858	588	.554698	.832
439	.425035	.904	489	.469743	.883	539	.513278	.858	589	.555530	.831
440	.425939	.905	490	.470626	.882	540	.514136	.857	590	.556361	.831
441	.426844	.904	491	.471508	.882	541	.514993	.857	591	.557192	.830
442	.427748	.904	492	.472390	.881	542	.515850	.857	592	.558022	.829
443	.428652	.903	493	.473271	.880	543	.516707	.856	593	.558851	.829
444	.429555	.903	494	.474151	.881	544	.517563	.855	594	.559680	.829
445	.430458	.902	495	.475032	.879	545	.518418	.855	595	.560509	.828
446	.431360	.902	496	.475911	.880	546	.519273	.854	596	.561337	.827
447	.432262	.902	497	.476791	.878	547	.520127	.854	597	.562164	.827
448	.433164	.901	498	.477669	.879	548	.520981	.853	598	.562991	.826
449	.434065	.901	499	.478548	.878	549	.521834	.853	599	.563817	.825
450	.434966	.900	500	.479426	.877	550	.522687	.852	600	.564642	.826

Table 2 - Angle - Sine Conversion Table (Sheet 3 of 5)

Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant
601	.565468	.824	651	.605982	.795	701	.644982	.764	751	.682370	.731
602	.566292	.824	652	.606777	.795	702	.645746	.763	752	.683101	.730
603	.567116	.823	653	.607572	.794	703	.646509	.763	753	.683831	.729
604	.567939	.823	654	.608366	.793	704	.647272	.762	754	.684560	.729
605	.568762	.822	655	.609159	.793	705	.648034	.761	755	.685289	.728
606	.569584	.822	656	.609952	.792	706	.648795	.761	756	.686017	.727
607	.570406	.821	657	.610744	.792	707	.649556	.760	757	.686744	.726
608	.571227	.821	658	.611536	.791	708	.650316	.759	758	.687470	.726
609	.572048	.819	659	.612327	.790	709	.651075	.759	759	.688196	.725
610	.572867	.820	660	.613117	.790	710	.651834	.758	760	.688921	.725
611	.573687	.819	661	.613907	.789	711	.652592	.757	761	.689646	.724
612	.574506	.818	662	.614696	.788	712	.653349	.757	762	.690370	.723
613	.575324	.817	663	.615484	.788	713	.654106	.756	763	.691093	.722
614	.576141	.818	664	.616272	.787	714	.654862	.755	764	.691815	.722
615	.576959	.816	665	.617059	.787	715	.655617	.755	765	.692537	.721
616	.577775	.816	666	.617846	.786	716	.656372	.754	766	.693258	.720
617	.578591	.815	667	.618632	.785	717	.657126	.754	767	.693978	.720
618	.579406	.815	668	.619417	.785	718	.657880	.753	768	.694698	.719
619	.580221	.814	669	.620202	.784	719	.658633	.752	769	.695417	.718
620	.581035	.814	670	.620986	.784	720	.659385	.751	770	.696135	.718
621	.581849	.813	671	.621770	.782	721	.660136	.751	771	.696853	.717
622	.582662	.812	672	.622552	.783	722	.660887	.750	772	.697570	.716
623	.583474	.812	673	.623335	.781	723	.661637	.750	773	.698286	.715
624	.584286	.811	674	.624116	.781	724	.662387	.748	774	.699001	.715
625	.585097	.811	675	.624897	.781	725	.663135	.749	775	.699716	.714
626	.585908	.810	676	.625678	.779	726	.663884	.747	776	.700430	.714
627	.586718	.810	677	.626457	.780	727	.664631	.747	777	.701144	.712
628	.587528	.808	678	.627237	.778	728	.665378	.746	778	.701856	.712
629	.588336	.809	679	.628015	.778	729	.666124	.746	779	.702568	.711
630	.589145	.807	680	.628793	.777	730	.666870	.744	780	.703279	.711
631	.589952	.808	681	.629570	.777	731	.667614	.745	781	.703990	.710
632	.590760	.806	682	.630347	.776	732	.668359	.743	782	.704700	.709
633	.591566	.806	383	.631123	.775	733	.669102	.743	783	.705409	.708
634	.592372	.806	684	.631898	.775	734	.669845	.742	784	.706117	.708
635	.593178	.804	685	.632673	.774	735	.670587	.742	785	.706825	.707
636	.593982	.804	686	.633447	.774	736	.671329	.740	786	.707532	.707
637	.594786	.804	687	.634221	.773	737	.672069	.741	787	.708239	.705
638	.595590	.803	688	.634994	.772	738	.672810	.739	788	.708944	.705
639	.596393	.802	689	.635766	.771	739	.673549	.739	789	.709649	.704
640	.597195	.802	690	.636537	.771	740	.674288	.738	790	.710353	.704
641	.597997	.801	691	.637308	.770	741	.675026	.738	791	.711057	.703
642	.598798	.801	692	.638078	.770	742	.675764	.736	792	.711760	.702
643	.599599	.800	693	.638848	.769	743	.676500	.736	793	.712462	.701
644	.600399	.799	694	.639617	.768	744	.677236	.736	794	.713163	.701
645	.601198	.799	695	.640385	.768	745	.677972	.735	795	.713864	.700
646	.601997	.798	696	.641153	.767	746	.678707	.733	796	.714564	.699
647	.602795	.798	697	.641920	.767	747	.679440	.734	797	.715263	.698
648	.603593	.797	698	.642687	.766	748	.680174	.733	798	.715961	.698
649	.604390	.796	699	.643453	.765	749	.680907	.732	799	.716659	.697
650	.605186	.796	700	.644218	.764	750	.681639	.731	800	.717356	.696

Table 2 - Angle - Sine Conversion Table (Sheet 4 of 5)

Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant	Angle	Sine	Interp. Constant
801	.718052	.696	851	.751940	.659	901	.783948	.621	951	.813997	.580
802	.718748	.695	852	.752599	.658	902	.784569	.619	952	.814577	.580
803	.719443	.694	853	.753257	.657	903	.785188	.619	953	.815157	.579
804	.720137	.694	854	.753914	.657	904	.785807	.618	954	.815736	.578
805	.720831	.692	855	.754571	.656	905	.786425	.617	955	.816314	.577
806	.721523	.692	856	.755227	.655	906	.787042	.617	956	.816891	.576
807	.722215	.692	857	.755882	.654	907	.787659	.616	957	.817467	.576
808	.722907	.690	858	.756536	.654	908	.788275	.615	958	.818043	.575
809	.723597	.690	859	.757190	.653	909	.788890	.614	959	.818618	.574
810	.724287	.689	860	.757843	.652	910	.789504	.613	960	.819192	.573
811	.724976	.689	861	.758495	.651	911	.790117	.613	961	.819765	.572
812	.725665	.687	862	.759146	.650	912	.790730	.611	962	.820337	.571
813	.726352	.687	863	.759796	.650	913	.791341	.611	963	.820908	.571
814	.727039	.687	864	.760446	.649	914	.791952	.611	964	.821479	.570
815	.727726	.685	865	.761095	.649	915	.792563	.609	965	.822049	.569
816	.728411	.685	866	.761744	.647	916	.793172	.609	966	.822618	.568
817	.729096	.684	867	.762391	.647	917	.793781	.607	967	.823186	.567
818	.729780	.683	868	.763038	.646	918	.794388	.607	968	.823753	.567
819	.730463	.683	869	.763684	.645	919	.794995	.607	969	.824320	.566
820	.731146	.682	870	.764329	.644	920	.795602	.605	970	.824886	.565
821	.731828	.681	871	.764973	.644	921	.796207	.605	971	.825451	.564
822	.732509	.680	872	.765617	.643	922	.796812	.604	972	.826015	.563
823	.733189	.680	873	.766260	.642	923	.797416	.603	973	.826578	.562
824	.733869	.679	874	.766902	.642	924	.798019	.602	974	.827140	.562
825	.734548	.678	875	.767544	.640	925	.798621	.601	975	.827702	.561
826	.735226	.677	876	.768184	.640	926	.799222	.601	976	.828263	.560
827	.735903	.677	877	.768824	.639	927	.799823	.600	977	.828823	.559
828	.736580	.676	878	.769463	.638	928	.800423	.599	978	.829382	.558
829	.737256	.675	879	.770101	.638	929	.801022	.598	979	.829940	.557
830	.737931	.675	880	.770739	.637	930	.801620	.597	980	.830497	.557
831	.738606	.674	881	.771376	.636	931	.802217	.597	981	.831054	.556
832	.739280	.673	882	.772012	.635	932	.802814	.596	982	.831610	.555
833	.739953	.672	883	.772647	.634	933	.803410	.595	983	.832165	.554
834	.740625	.672	884	.773281	.634	934	.804005	.594	984	.832719	.553
835	.741297	.670	885	.773915	.633	935	.804599	.593	985	.833272	.553
836	.741967	.670	886	.774548	.632	936	.805192	.593	986	.833825	.551
837	.742637	.670	887	.775180	.631	937	.805785	.592	987	.834376	.551
838	.743307	.668	888	.775811	.631	938	.806377	.591	988	.834927	.550
839	.743975	.668	889	.776442	.630	939	.806968	.590	989	.835477	.549
840	.744643	.667	890	.777072	.629	940	.807558	.589	990	.836026	.548
841	.745310	.667	891	.777701	.628	941	.808147	.589	991	.836574	.548
842	.745977	.665	892	.778329	.627	942	.808736	.588	992	.837122	.546
843	.746642	.665	893	.778956	.627	943	.809324	.587	993	.837668	.546
844	.747307	.664	894	.779583	.626	944	.809911	.586	994	.838214	.545
845	.747971	.663	895	.780209	.625	945	.810497	.585	995	.838759	.544
846	.748634	.663	896	.780834	.625	946	.811082	.585	996	.839303	.543
847	.749297	.662	897	.781459	.623	947	.811667	.584	997	.839846	.543
848	.749959	.661	898	.782082	.623	948	.812251	.582	998	.840389	.541
849	.750620	.660	899	.782705	.622	949	.812833	.583	999	.840930	.541
850	.751280	.660	900	.783327	.621	950	.813416	.581	1.000	.841471	.540

Table 2 - Angle - Sine Conversion Table (Sheet 5 of 5)

INDEX

Abbe condenser 23.3.3.1.1
Abbe constant 2.7.2
Abbe prism, type A 13.10.4
Abbe prism, type B 13.10.5
Aberration 5.11.3.2
Aberration, chromatic 6.10.1
Aberration coefficients, first order stop shift 8.9.4
Aberration coefficients, spherical...8.5.1.1
Aberration coeffieient, third order ...8.5.2
Aberration coefficients, third order stop shift 8.9.3
Aberration, fifth order 5.11.3.3
Aberration, first order 5.11.3.4
Aberration of prisms, third order ... 13.9.2
Aberration,spherical 5.10.2.2
Aberrations 24.2.2
Aberrations, oblique 8.9.1
Absentee (non - absorbing) layers . 21.2.14.1
Absolute refractive index 2.5.1
Absorption 17.2.3
Absorption coefficient 21.2.2
Accomodation, eye 4.4.4
Achromatic beam splitters.. 20.1.2.2, 20.7.2
Achromatic microscope objective ... 23.3.4.1
Achromatic system, thin lens 6.10.8
Achromatization 21.10.7
Achromatized bilayers 21.7.4
Acuity, visual (VA) 4.5.4
Admittances, method of 21.2.12
Aerosol 18.2.2.6
Afocal 6.3.7.1
Age, eye 4.8.2
Air equivalent prism 13.8.3.1
Air space, use of 19.2.3
Airy disc 3.1.1.1
Albinism 4.2.1
Amblyopia 4.1.3
Ametropia 4.1.3
Amici prism 13.10.7
Amplitude reflectance and transmittance Fresnel coefficients 16.8.1
Analogics 20.1.4
Analogics, stack 20.4.5
Aniseikonia 4.7.4
Anisotropy 17.3.4
Angle, critical 2.4.1
Angle shift, minimization of 20.4.7
Angle sign convention 2.2.2
Angles, incidence, reflection refraction 2.1.4
Angles; slope 5.6.5.1
Angles, visual 4.4.4.5
Angular aberrations of telescopes . 11.1.3.5
Angular chromatic aberrations 7.3.7.1
Angular magnification 6.3.7.2
Ann Arbor tester 25.9.3
Annealed, fine 2.7.8
Antireflection coatings 20.1.2.1
Aperture , effective system 22.6.6.9
Aperture , numerical (NA) 7.3.3, 23.2.5.2
Aperture stop 6.11.2
Aperture stop position, periscope ... 7.5.5
Aperture stop and pupils, microscope..7.3.4
Aphakic 4.4.4.2
Aplanatic condenser 23.4.4.2
Apochromatic microscope objective ..23.3.4.1
Apparent field 7.6.2.1
Apparent prism thickness 13.8.31
Approximation, first order 5.9.2.2
Approximation, zeroth 5.11.1.4
Areas, Panum's 4.7.2.2
Aspheric dark field condenser 23.12
Aspheric dark field condensers 23.4.4
Aspheric effects, fourth order 8.7.3
Aspheric surface, mathematical description of 5.5.2
Aspherica 24.3.2
Asthenopia 4.8.1
Astigmatic constant 2.2.4.4
Astigmatism 4.4.1.2
Astigmatism and curvature of field test 25.5
Astigmatism - with the rule, against the rule 4.4.13
Astronomical extinction 18.2.1.1
Atmospheric contaminants, sources and effects 18.7.1
Auxiliary optical measurements 25.7
Axial and lateral color, graphical interpretation of 6.10 5
Axial chromatic aberrations 6.10.1
Axial color 6.10.1
Axial color, correction of 6.10.6.1
Axial paraxial ray 6.3.1
Axial ray 5.9.2.2
Axiom, two mirror image location...... 13.6.2

Back focal length 8.3.2
Background 22.6.5.2
Bands, shadow 18.5.3.1
Band pass filter 20.9
Barr and Strand ocular prism 13.10.19
Basic lens designing procedure, analysis of 9.2
Basic optical system 5.1.1
Beam splitter 20.7.1
Beam splitters, achromatic...20.1.2.2, 20.7.2
Beam splitters, color selective 20.1.2.4
Bending, lens 9.2.4.10
Berthele eyepiece 14.8
Bilayer coatings 21.7
Bilayer coatings, methods of computation 21.7.2
Bilayer coatings, simplest form 21.7.3
Bilayers, achromatized 21.7.4
Bilayers, non - quarter wave 21.7.6
Bilayers, null - isoachromatic 21.7.4.1
Bilayers, quarter wave 21.7.5
Binocular design considerations 4.7.5
Binocular design limitations 4.7.4
Blocking filter 20.10.1.2
Box in 10.2.2.5
Broad - band reflectors 20.5.1.3.1
Brightfield microscopy 23.4.1
Bubbles 2.7.7

Calculation of aberration, ray trace procedure 6.2.6
Camera, satellite tracking 19.5.1
Cardinal points 6.5.1.1
Carl Zeiss binocular - ocular prism system 13.10.21
Carl Zeiss coincidence prism system 13.10.20

Carl Zeiss ocular prism 13.10.18
Carl Zeiss prism system 13.10.16
Cassegrain telescope system 19.3.3
Catadioptric systems 19.4
Cataract 4.4.4.2
Centered optical system 5.1.2.1
Channeled spectra 16.19.1.1
Characteristics of the human eye 4.1.1
Characteristics, infrared detector 22.6.4
Chart, Kinetic definition 26.1.4
Chief ray 6.3.1, 6.11.4, 8.3.1.1
Chief ray, other paraxial 8.9.2.1
Chief ray, shifted 8.9.2.1
Chromatic aberration 6.10.1
Chromatic aberration, angular 7.3.7.1
Chromatic aberration, basic concepts
  in correcting for 6.10.6
Chromatic aberration, lateral 6.10.1
Chromatic aberration, longitudinal 6.10.1
Chromatic aberration,
  longitudinal axial 6.10.5.1
Chromatic aberration, particular
  wavelengths for calculation of 6.10.4
Chromatic aberration, simple
  afocal microscope 7.3.7
Chromatic aberration, thin lens 6.10.7.1
Chromatic aberration, total 6.10.3
Chromatic aberration, transverse 6.10.1
Chromatic aberration, transverse (Tch)..6.10.3
Chromatic aberration, transverse
  axial (TAch) 6.10.3
Chromatic abberations, axial 6.10.1
Chromatic distribution formulae 9.3.2
Chromatic surface coefficients 6.10.3
Chromatic variation of
  spherical aberration 11.3.2.4
Circular apertures, resolution with 16.27
Clark lens 19.2.5.4
Close ray 5.8.1.1
Clouding effect 20.2.4.2.5
Coating, double-quarter
  double minimum 20.3.4.2.4
Coating, double-quarter
  single minimum 20.3.4.2.2
Coating, antireflection 20.1.2.1
Coating, bilayer 21.7
Coating, double layer 20.3.4
Coating, monolayer 21.6
Coating, simple metal 17.7.3
Coating, single layer 20.3.3
Coating, triple layer antireflection .. 20.3.5
Coddington equations 5.8.4
Coefficient, absorption or
  extinction 21.2.2
Coefficient, differential 6.9.2.2
Coefficient, reflection 17.2.2
Cold mirror 20.1.2.7, 20.5.3
Collinear, coherent waves 3.3.1
Collinear, incoherent waves 3.3.2
Color and band pass filters 20.1.2.3
Color, axial 6.10.1
Color filters, long wave pass 20.5.2
Color lateral 6.10.1
Color-selective beam splitters 20.1.2.4
Color, secondary 6.10.8.4
Color vision 4.5.6
Colors, interference 23.7.4
Coltman variable frequency
  square wave test 26.4.2
Coma correction 11.3.6
Compensating eyepiece 23.3.5.4
Complex numbers, physical optics 16.14
Complex reflectance, approximate
  method of computation 21.2.8.6
Complex transmittance 21.2.8.7
Composite filter 20.5.1.3, 20.5.1.3.2
Compound microscope 23.1.2.9
Concept, thin lens 6.7.1
Condenser, Abbe 23.3.3.1.1
Condenser, aplanatic 23.4.4.2
Condenser, epi- 23.3.2.4.2
Condenser, substage 23-2.5.1
Condensers, substage 23.3.3
Condition, quarter wave............... 21.3.5
Condition, zero 21.2.16
Conjugates, finite 6.6.2
Conjugates, infinite 6.6.1
Conjugate planes 6.5.7
Constant, Abbe 2.7.2
Constant, astigmatic.................. 2.2.4.4
Constants. optical 21.2.1
Constraint, differential
  equation of 5.8.2.3
Constraint, equation of 5.8.2.3
Contrast 4.5.4.2
Contrast and time 4.5.2
Contrast, densiphase 23.6.5
Contributions, surface 6.10.1
Cornea 4.4.1
Corrected lens 6.10.5
Correction of the axial color 6.10.6.1
Critical angle 2.4.1
Critical angles and indices,
  table of 2.4.2
Critical flicker frequency (CFF) 4.5.3
Critical illumination 23.3.2.1
Crystals, optical 17.6.2.2
Curl relation, Maxwell's 21.2.1
Currents, dome 18.6.3
Current, tube 18.6.2
Curvature of a field 6.10.8.5
Curvature, Petzval 6.10.8.5
Curves , field 8.6.2.2
Cutoff 20.5.1.2
Cutoff filter 20.4.7.2
Cutoff wavelength 20.5.1.2

Darkfield microscopy 23.4.1
Darkfield condensers, aspheric 23.4.4
Darkfield condensers, reflecting ... 23.4.3
Darkfield condensers, refracting ... 23.4.2
Darkfield condensers, spherical 23.4.5
Data, initial ray 5.4.6.1
Data system, initial 5.4.6.1
Deformation terms 5.5.2.3
Definitions and conventions 5.2.2
Densiphase contrast 23.6.5
Design phase testing 24.1.4.2
Designing a lens system,
  approach to 9.1.2
Deutan 4.5.6.2
Deuteranomalous 4.5.6.1
Deuteranopes 4.5.6.1
Development of the eye 4.2.4
Diagram, spot 8.2.1
Diagram, tunnel 13.8.1.1
Dialyte 6.7.3.1
Dichroic mirror 20.1.2.4
Dichroic mirrors............ 20.7.1, 20.7.3
Dichroism 20.7.3
Dichromatic vision 4.5.6.1

Differential coefficients 6.9.2.2
Differential equation of constraint...5.8.2.3
Differentially traced ray 5.8.1.1
Differentially traced meridional ray 5.8.3
Differentially traced skew ray 5.8.2
Differential refraction equations ... 5.8.2.2
Differential transfer equations 5.8.2.2
Diffraction 3.1.2.1
Diffraction, Fresnel 16.22.1.6
Diffraction from spherical wave fronts 16.25
Diffraction image 3.1.1.1
Diffraction nature of optical images ...3.1.1
Diffraction plate 23.6.4
Diffusion 17.2.4
Dimming, surface 2.7.6
Diopters 11.1.3.5
Direction cosines, optical 5.2.2
Direction of rays 2.1.3
Disc, Airy 3.1.1.1
Disc, seeing 18.5.3.2
Dispersion 2.6.2.2
Dispersion, mean 2.7.3
Dispersion, partial 2.7.3
Displacement, transverse 6.10.3
Distance, interpupillary 4.7.2
Distortion 25.6.3, 8.6.1.3
Distortion, fractional 8.6.1.3
Distortion test 25.6
Distribution curve, energy 8.2.3.1
Dome currents 18.6.3
Double dove prism 13.10.11
Double half-wave system 20.10.7
Double layer coatings 20.3.4
Double-quarter double minimum coating 20.3.4.2.4
Double -quarter single ~~minimum~~ coating 20.3.4.2.2
Double relay systems 12.6
Doublets as relay lenses 12.5
Doublet, telescope 11.2.1.1
Doublet, thick lens 11.3.1
Dynamic visual acuity,(DVA) 4.5.4.2

Effect, Stiles- Crawford 4.4.5.1
Effective aperture function 22.6.6.9
Effective interface 20.10.2.1
Effective system aperture 22.6.6.9
Effects, polarization 20.7.1.2
Efficiency , KDC 26.1.4.4
Electric and magnetic vectors 3.2.1.1
Electromagnetic waves, velocity of .. 3.2.1.2
Emmetropia 4.1.3
Empty field myopia 4.4.4.1
Energy density, time averaged 3.1.2.2
Energy density, time averaged 16.1.1.3
Energy distribution curve............ 8.2.3.1
Energy in a single wave 3.2.3
Energy, instantaneous 3.2.3
Energy reflectance 21.2.5
Energy transmittance 21.2.5.6
Ensemble multilayer 20.5.1.3
Entrance pupil...................... 6.11.3.1
Entrance pupil plane 6.11.3.1
Entrance pupil point 6.11.3.1
Entrance pupil, ray distribution in ... 8.2.2
Epi-condensers 23.3.2.4.2
Equal inclination 16.11.1.5
Equation of constraint 5.8.2.3
Equations, Coddington 5,8,4

Equations, differential refraction....5.8.2.2
Equations, differential transfer 5.8.2.2
Equations, paraxial ray trace 9.3.1
Equations, refraction 5.3.2
Equations, stop shift 9.3.4, 8.9.5.1
Equations, transfer 5.3.1
Erect left-handed image...............7.5.1.3
Erfle eyepiece 14.9
Erfle eyepieces, modified 14.10
Esophoria 4.6.4
Esotropia 4.6.4
Evaluation, image 9.2.8
Evaluation phase testing 24.1.4.4
Exit pupil 6.11.3.2
Exit pupil plane.................... 6.11.3.2
Expansion of the optical sine function 5.11.1
Extinction, astronomical 18.2.1.1
Extinction coefficient 21.2.2
Extinction, photographic instruments ... 18.4
Extinction, visual instruments 18.3
Eye, development of 4.2.4
Eye, physical structure of 4.2.1
Eye accomodation....................... 4.4.4
Eye age 4.8.2
Eye fatigue 4.8.1
Eye relief 7.3.8.2
Eye resolution 4.4.5
Eye sensitivity 4.5.1
Eyepiece, basic functions 14.1.1
Eyepiece, Berthele 14.8
Eyepiece, design considerations 14.1.2
Eyepiece, Erfle 14.9
Eyepiece, high-eyepoint 23.3.5.4
Eyepiece, Huygenian .....14.3.........23.3.5.2
Eyepiece, Kellner 14.5
Eyepiece, modified Erfle 14.10
Eyepiece, orthoscopic 14.6
Eyepiece, Ramsden 23.3.5.3, 14.4
Eyepiece, symmetrical (Plössl) 14.7
Eyepiece, Wild 14.11
Eyepieces, compensating 23.3.5.4
Eyepieces, microscope 23.3.5

Fabry-Perot all dielectric filters .. 20.10.4
Fabry-Perot all dielectric filters for the infrared 20.10.6
Fabry-Perot all dielectric filters for the visible 20.10.5
Fabry-Perot filters 20.9
Fabry-Perot filters with non-Lorentzian shaped band pass 20.10.7
Fabry-Perot interferometer, basic concepts 20.10.1
Fabry-Perot multilayer filters 20.10.2
Far infrared region image quality 22.6.2
Far point 7.2.5
Far-sightedness 4.3.3.1
Fatigue, eye 4.8.1
Fat lens 7.3.6
Field, apparent 7.6.2.1
Field angle, half image 7.2.2
Field angle, half object 7.2.2
Field, curvature of 6.10.8.5
Field curves 8.6.2.2
Field flat 6.10.8.5
Field lens 7.3.8.1
Field lens, effects of 7.3.8
Field lenses, periscopes 7.5.4
Field of view 7.3.1, 22.6.3.2
Field, real 7.6.2.1

Field, stop 7.3.1
Fifth order aberration 5.11.3.3
Films, thin 17.7.1
Filter, band pass 20.9
Filter, blocking 20.10.1.2
Filter, composite 20.5.1.3, 20.5.1.3.2
Filter, cutoff 20.4.7.2
Filter, multilayer 20.1.1
Filters 18.3.2
Filters, color and band pass 20.1.2.3
Filters, Fabry-Perot 20.9
Filters, Fabry-Perot all dielectric 20.10.4
Filters, Fabry-Perot multilayer 20.10.2
Filters, heat control 20.1.2.7
Filters, infrared long-wave pass 20.5.4
Filters, interference 20.1.2.5
Filters, metal film band pass 20.10.3
Filters, narrow band pass 21.10.8
Filters, reflection 20.1.2.10
Filters, short-wave pass 20.6
Filters, short-wave pass color 20.6.1
Filters, short-wave pass infrared 20.6.3
Filtering, spatial 22.6.5.3
Fine annealed 2.7.8
Final angle, effect of curvature change on 6.9.4
Final angle, effect of thickness change on 6.9.5
Finite angles and heights for paraxial rays, use of 5.9.3
Finite conjugates 6.6.2
First focal length 6.5.5.3
First focal plane 6.5.5.2
First focal point 6.5.5.2
First order aberration 5.11.3.4
First order approximation 5.9.2.2
First order imagery in a mirror 6.8.3
First order optical system 8.1.1
First order optics 5.9.22, 5.11.2
First order quantities 5.8.1.1
First order thin lens 11.2.1
First principal plane 6.5.5.3
First principal point 6.5.5.3
First, second, third...etc orders 5.11.1.4
First surface reflection 17.7.2
Fixed frequency square wave test 26.4.3
Fizeau fringes 16.12.1.1
Fizeau interferoscope 25.8.2
Fizeau interferoscope, principles of operation 16.2.1
Fizeau interferoscope, testing for optical flatness with 16.2.2
Flat field 6.10.8.5
Flicker 4.5.3
Flicker frequency, critical (CFF) 4.5.3
- number 7.3.5
Focal length, back 8.3.2
Focal length, effect of curvature change on 6.9.3
Focal length, first 6.5.5.3
Focal length, second 6.5.2
Focal length test 25.2
Focal plane, first 6.5.5.2
Focal plane, second 6.5.2
Focal point, first 6.5.5.2
Focal point, image 6.5.2
Focal point, second 6.5.2
Focus, sagittal 5.8.4.1
Focus, skew 5.8.4.1
Focus, tangential 5.8.4.3
Form, Gaussian 6.5.6.4
Form, Newtonian 6.5.6.3

Format, paraxial ray trace 6.2.2
Formula, interference 3.3.3.3
Formulae, chromatic distribution 9.3.2
Form, short 5.6.3.2
Foucault test 25.10
Fourth order aspheric effects 8.7.3
Fractional distortion 8.6.1.3
Frankford Arsenal prism No.1 .... 13.10.22
Frankford Arsenal prism No.2 .... 13.10.23
Frankford Arsenal prism No.3 .... 13.10.24
Frankford Arsenal prism No.4 .... 13.10.25
Frankford Arsenal prism No.5 .... 13.10.26
Frankford Arsenal prism No.6 .... 13.10.27
Frankford Arsenal prism No.7 .... 13.10.28
Fraunhofer diffraction, discussion of 16.22.1
Fraunhofer diffraction from circular apertures, discussion of 16.24.1
Fraunhofer diffraction from rectangular apertures, discussion of 16.23.1
Fraunhofer lines 2.6.3
Fraunhofer objectives 11.2.2.3
Free spectral range 20.10.1.2
Frequency 3.2.1.2
Fresnel's coefficients, summary of 21.2.7
Fresnel diffraction 16.22.1.6
Fringe width 3.3.3.3
Fringes, Fizeau 16.12.1.1
Function, pupil 16.28.1.2
Function, merit 9.2.7.3
Funnel stop 23.4.3

Galilean telescope, analysis of 7.6.2
Gaussian form 6.5.6.4
General p:g stack 20.4.4
General ray 5.4.1.1
Geometrical optics 2.1.1
Glare 4.5.5.2
Glass for infrared usage 22.2.3
Glass types, lenses 10.2.2
Glass type number 2.7.4
Glaucoma 4.2.3.1
Goerz prism system 13.10.17
Graphical ray tracing, explanation of 5.7.1

Haidinger's interference fringes, interpretation of 16.11.1
Half image field angle 7.2.2
Half object field angle 7.2.2
H and L layers 20.1.3.5
Harting - Dove prism 13.10.10
+ Heat control filters 20.1.2.7
Heat haze, Summer 18.2.2.9
Heat reflector 20.1.2.7
Heat reflectors 20.6.2
Herpin equivalent index 20.4.8.3.1
Herriot electronic lens bench 26.3.3
Heterophoria 4.6.4
Heterotropia 4.6.4
High-eyepoint eyepiece 23.3.5.4
High index substrate 20.3.4.5.1

+ HARTMAN 25.3.2

High magnification 23.2.2
High reflectance zone 20.4.1.2
High reflectivity mirrors 20.1.2.8
High resolution, microscope 23.2.5
Homogeneity, optical 17.3.2
Hot mirror 20.6.2.1
Human eye, characteristics of 4.1.1
Humor, vitreous 4.4.4.4
Huygenian eyepiece 23.3.5.2, 14.3
Huygens' principle 16.21
Hyperopia 4.3.3.1
Hyper-hypostereoscopy 4.7.5

Illumination, critical 23.3.2.1
Illumination, Kohler 23.3.2.2
Illumination, microscope 23.2.4
Illumination, variation of 25.10.2.2
Illumination, vertical 23.2.4
Illumination systems, microscope 23.3.2
Illuminator, optical requirements 23.3.2.3
Image converter systems, infrared 22.5.2
Image converter tube 22.2.1
Image erection by lenses 7.5.3
Image evaluation 9.2.8
Image focal point 6.5.2
Image forming device 22.4.2
Image inversion, microscopes and telescopes 7.5.2
Image location, mirror 13.3.2.1
Image location, two mirror 13.6.1
Image, diffraction 3.1.1.1
Image orientation, periscope 7.5.1
Image, perverted 7.5.1.3
Image motion 18.5.1.5
Image plane, ray distribution in 8.2.3
Image quality, far infrared region 22.6.2
Image quality, intermediate infrared region 22.6.2
Imagery, infrared 22.2.2
Image sphere, mirror 13.5.1.1
Image, 0° polynomial 8.5.1
Imbalance 4.6.3
Immersed 22.6.7.7
Incidence, non-normal 20.1.3.7
Incidence, plane of 13.3.1
Inclination, equal 16.11.1.5
Inclusions 2.7.7
Index, Herpin equivalent 20.4.8.3.1
Index, refractive 3.2.1.2
Index of refraction 2.2.1
Index of refraction, absolute 2.5.1
Index of refraction, relative 2.5.2
Indices, reference 2.7.1
Inferior oblique muscles 4.6.2
Inferior erectus muscle 4.6.2
Infinite conjugate 6.6.1
Infrared 2.6.1
Infrared absorption, optical glass 22.2.4
Infrared applications 22.3.1
Infrared detector characteristics 22.6.4
Infrared image converter systems 22.5.3
Infrared imagery 22.2.2
Infrared long-wave pass filters 20.5.4
Infrared materials, choice of 22.3.2
Infrared material, size limitations of 22.3.3
Infrared optical system, general functions of 22.6.3
Infrared photography 22.5.2
Infrared short-wave pass filters 20.6.3
Infrared system, triggered radiation type 22.5.4
Infrared target detection and location 22.6.5
Infrared wavelength range 22.4.3
Initial data, skew ray 5.4.2
Initial ray data 5.4.6.1
Initial system data 5.4.6.1
Instantaneous energy 3.2.3
Instrument orientation, effect of 18.4.1
Interference 3.1.2.2
Interference colors 23.7.4
Interference formula 3.3.3.3, 16.1.1.5
Interference filter 20.9
Interference filters 20.1.2.5
Interference path 21.3.5
Interference with plane parallel plates and distant light sources, discussion of 16.9.1
Interference with plane parallel plates and nearby light sources, discussion of 16.10.1
Interferometer, Lloyd's 16.7
Interferometer, modified Michelson 25.8.3
Interferometer, Twyman-Green 25.8.4.4
Interferoscope, Fizeau 25.8.2
Intermediate infrared region image quality 22.6.2
Internal transmittance 17.2.3.2
Interpupillary distance 4.7.2, 13.10.21
Intraocular pressure 4.2.2
Invariant, optical 6.3.2
Invariant position of the two mirror image 13.6.3
Invert 13.10.1.2
Inverted right-handed image 7.5.1.2
Isoachromatic 21.7.4.1
Isotropy, optical 17.3.4
Iteration procedure 5.5.4.6
Inward curving field 10.2.2.4

K.D.C. efficiency 26.1.4.4
Kellner eyepiece 14.5
Keratoconus 4.4.1.2
Kinetic definition chart 26.1.4
Kohler illumination 23.3.2.2

Lagrange equations 6.3.8
Lambert system testing 26.3.2
Landolt C 4.5.4
Laser 20.8.2
Lateral chromatic aberration 6.10.1
Lateral color 6.10.1
Lateral magnification 6.3.6
Lateral rectus muscle 4.6.2
Law of reflection 2.3.2
Law of refraction 2.2.3
Law of refraction, vector form 2.2.4
Law, Talbot's 4.5.3
Layers, absentee (non-absorbing) 21.2.14.1
Layers, H and L 20.1.3.5
Layers, matched 20.1.6.3
Left-handed image 7.5.1.3
Left-hand triplet solution 10.3.2.2
Leman prism 13.10.6
Lens, Clark 19.2.5.4
Lens, corrected 6.10.5
Lens, eye 4.4.3

Lens, fat ... 7.3.6
Lens, field ... 7.3.8.1
Lens, over-corrected ... 6.10.5.1
Lens, relay ... 7.3.2
Lens, single ... 7.2.1
Lens, Taylor triplet ... 10.1.1
Lens, glass types ... 10.2.2
Lens power and spacing, example of .. 10.2.1
Lens problem of a relay system ... 12.2
Lens relay system ... 12.1
Lens thickness ... 10.4.1
Lens, thin ... 6.7.1.2
Lens, undercorrected ... 6.10.5
Lens bending ... 9.2.4.10
Light ... 4.1.1
Light, F,C, and D, difference in focus of ... 11.4.1
Lighting ... 4.5.5
Light gathering power ... 18.3.3
Light wave plane polarized ... 3.2.2.1
Limit of resolution ... 23.2.5.2
Lines, Fraunhofer ... 2.6.3
Lloyd's interferometer ... 16.7
Long to short conjugate ... 14.2.1
Long wave pass color filters ... 20.5.2
Long wave pass filters, general properties of ... 20.5.1
Longitudinal axial chromatic aberration ... 6.10.5.1
Longitudinal chromatic aberration ... 6.10.1
Longitudinal spherical aberration .. 8.5.3.2
Longitudinal spherical aberration test ... 25.3
Loss of vision ... 4.1.3
Low index substrate ...20.3.4.3.1

Magnification, angular ... 6.3.7.2
Magnification, high ... 23.2.2
Magnification, lateral ... 6.3.6
Magnification, negative ... 7.5.1.2
Magnification, positive ... 7.5.1.1
Magnification, unit positive ... 6.5.7
Magnifying power (MP) ... 6.3.7.2, 7.2.2
Magnifying power, telescope ... 7.4.2
Magnifying power equation, analysis of ... 7.2.5
Manufacturing phase testing ... 24.1.4.3
Margin ... 6.11.3.1
Marginal ray ... 6.4.3.2
Matched layers ... 20.1.6.3
Mathematical description of an aspheric surface ... 5.5.2
Mathematics for mirror image location ... 13.4.2
Matrices and quarternions corresponding ... 21.5.3
Matrix ... 21.4.2.2
Matrix, square ... 21.4.2.3
Matrix form, vector ray tracing ... 13.4.3
Matrix methods ... 21.4
Maximum light-receiving ability .. 22.6.6.11
Maxwell's curl relation ... 21.2.1
Mean dispersion ... 2.7.3
Measurements with monochromatic light..16.18
Measuring vision ... 4.5.4
Mechanical strain ... 17.3.3
Medial rectus muscle ... 4.6.2
Meridional ray ... 5.4.1.1, 5.6.1
Meridional ray fan ... 8.3.1.1
Meridional ray trace, aspheric surfaces ... 5.6.4
Meridional ray trace, spherical sufaces ... 5.6.3
Meridional ray trace, spherical surfaces (simplified) ... 5.6.5
Merit function ... 9.2.7.3
Mesoptic vision ... 4.4.4.9
Metabolism ... 4.2.3
Metal film band pass filters ... 20.10.3
Method, refining ...20.4.8.3, 20.4.8.4
Methods, matrix ... 21.4
Methods, quaternion ... 21.5
Michelson interferometer, modified....25.8.3
Microscope, aperature stop & pupils .. 7.3.4
Microscope, compound ...23.1.2.9
Microscope components, functional relationships ... 23.1.2
Microscope eyepieces ... 23.3.5
Microscope eyepiece, design difficulties ... 7.3.6
Microscope illumination ...23.2.4
Microscope illumination systems ... 23.3.2
Microscope illumination and filters ..23.7.5
Microscope illuminator, vertical .. 23.3.2.4
Microscope objectives ... 23.3.4
Microscope objective, achromatic .. 23.3.4.1
Microscope objective, apochromatic ... 23.3.4.1
Microscope objective, semi-apochromatic ... 23.3.4.1
Microscope paraxial ray trace ... 7.3.3
Microscope, simple ... 7.3.2
Microscopy, bright field ... 23.4.1
Microscopy, dark field ... 23.4.1
Mirror, cold ... 20.1.2.7, 20.5.3
Mirror, dichroic ... 20.1.2.4
Mirror, hot ... 20.6.2.1
Mirror image location ... 13.3.2.1
Mirror image location, mathematics for ... 13.4.2
Mirror image sphere ... 13.5.1.1
Mirror imagery, single ... 13.4.1
Mirrors and prisms, use of ... 13.1.1
Mirrors, dichroic ...20.7.1, 20.7.3
Mirrors, high reflectivity ... 20.1.2.8
Mirrors, opaque ... 20.8.1
Mirrors, semi-transparent ..20.1.2.6, 20.8.2
Modified Erfle eyepiece ... 14.10
Monolayer coatings ... 21.6
Monolayer coatings, methods of computation ... 21.6.2
Motion, image ... 18.5.1.5
Multilayer ... 20.1.1
Multilayer filter ... 20.1.1
Multilayer filters, methods of deposition ... 20.2.2
Multilayer filters, substrates for .. 20.2.3
Multilayer stack ... 20.1.3.2
Multilayers, normal incidence upon .. 21.2.8
Multilayers, oblique incidence upon ... 21.2.9
Multilayers, quarter-wave ... 21.10
Multiple beam interference fringes from slightly inclined surfaces ... 16.17
Multiple reflection ... 13.2.2.1
Multiple scattering ... 18.2.2.10
Muscular action, eye ... 4.6.2
Mydriasis ... 4.4.2.2
Myopia ... 4.3.3.1
Myopia, empty field ... 4.4.4.1
Myopia, night ... 4.4.4.1
Myosis ... 4.4.2.2

Narrow band pass filters 21.10.8
National Bureau of Standards
resolving power test 26.1.2
Near infrared region 22.5.1
Near point 7.2.5
Near-sightedness 4.3.3.1
Negative eyepiece,
Gallilean telescope 7.6.1
Negative magnification 7.5.1.2
Newton's reflective system 19.3.2
Newton's fringes, interpretation of .. 16.13.1
Newtonian form 6.5.6.3
Night myopia 4.4.4.1
Night vision devices 22.2.1
Nodal points, additional
characteristics of 6.5.8
Nodal point, second 6.5.4
Non-absorbing monolayers and substrates,
zero reflectance from 21.3
Non-absorbing systems,
normal incidence 21.6.3
Non-absorbing systems,
oblique incidence 21.6.7
Non-collinear coherent waves 3.3.3
Non-image forming device 22.4.2
Non-normal incidence 20.1.3.7
Non-quarter wave bilayers 21.7.6
Normal incidence, Fresnel's
coefficient for 21.2.3
Nu value 2.7.2
Null-isoachromatic bilayers 21.7.4.1
Number, ƒ 7.3.5
Number, glass type 2.7.4
Number, vee 2.7.2
Number, wave 20.1.3.3
Numerical aperture NA 7.3.3, 23.2.5.2
Numerical apertures, zonal 16.26.1.3
Nystagmus, physiological 4.6.3

Object and image points w/respect to
focal and principal points 6.5.6
Objectives, Fraunhofer............. 11.2.2.3
Objectives, microscope 23.3.4
Objective, microscope, achromatic ..23.3.4.1
Objective, microscope, apochromatic.23.3.4.1
Objective, microscope.
semi-apochromatic 23.3.4.1
Objective and eyepiece design,
telescope 7.4.3
Oblate spheroid 19.4.2.6
Oblique aberrations 8.9.1
Oblique image polynomial 8.6.1
Oblique incidence,
Fresnal's coefficients for 21.2.4
Oblique paraxial ray 6.3.1
Offense against the
sine condition (OSC) 11.3.4.3
Opaque mirrors 20.8.1
Optical anisotropy 17.3.4
Optical constants,
physical significance 21.2.2
Optical constants, thin films 21.2.1
Optical crystals 17.6.2.2
Optical design,
methods and problems 24.1.3
Optical direction cosines 5.2.2
Optical devices, testing 25.8
Optical glass characteristics,
table of 2.7.8
Optical glass infrared absorption ... 22.2.4

Optical half-width 16.15.1.11
Optical homogeneity 17.3.2
Optical images,
diffraction nature of 3.1.1
Optical invariant 6.3.2
Optical isotropy 17.3.4
Optical materials for wavelengths
longer than 2.2 μ 17.6.3
Optical materials for wavelengths
shorter than 0.36μ 17.6.4
Optical materials, imperfections in .. 17.4.1
Optical measurements, auxiliary 25.7
Optical object-image relationship 24.2.1
Optical path......... 2.1.3, 21.2.8.7, 23.6.1
Optical path difference 16.9.1.1
Optical plastics 17.6.2.3
Optical system, basic 5.1.1
Optical system,
environmental requirements 17.5.1
Optical system, first order 8.1.1
Optical systems involving mirrors,
sign convention of 6.8.1
Optical system, pupils as surfaces in..6.11.6
Optical testing devices 25.8
Optical tube length 23.2.3
Optics, fifth order 5.11.3.3
Optics, first order 5.9.2.2, 5.11.2
Optics, geometrical 2.1.1
Optics, order of 5.11.1.1
Optics, paraxial ray 5.9.2.2
Optics, third order 5.11.3.1
Order of optics 5.11.1.1
Order, zeroth 5.11.1.4
Orders; first, second. etc............ 5.11.1.4
Orthophoria 4.6.4
Orthoscopic 4.7.5
Orthoscopic eyepiece 14.6
Orthotropia 4.6.4
Oscillation 18.5.2
Other paraxial chief ray 8.9.2.1
Over-corrected lens 6.10.5.2

Panum's areas 4.7.2.2
Paraxial ray 5.4.1.1, 5.9.1
Paraxial ray optics 5.9.2.2
Paraxial ray trace format 6.2.2
Paraxial ray trace,
algebraic example 6.2.3
Paraxial ray trace, microscope 7.3.3
Paraxial ray trace,
numerical example 6.2.4
Paraxial ray tracing, importance of ... 6.2.1
Paraxial ray, axial 6.3.1
Paraxial ray, oblique 6.3.1
Paraxial ray trace equations 9.3.1
Paraxial ray tracing equations 5.9.2
Parfocal............................ 23.3.4.2
Partial dispersion 2.7.3
Partial dispersion ratio................ 2.7.3
Particulate light scattering 18.2.2
Path, interference 21.3.5
Path length, optical 2.1.3
Path, optical 2.1.3, 21.2.8.7, 23.6.1
Pechan prism 13.10.12
Penta prism 13.10.14
Perceptim 4.5.7
Period of vibration 3.2.1.2
Periodic structure,
variation on the 20.4.8
Periscope, field lenses 7.5.4

Periscope aperture stop position 7.5.5
Perverted image 7.5.1.3
Petzval curvature 6.10.8.5
Phase change on reflection 16.8.1.2
Phase retardation 21.6.2.2
Phase velocity 3.2.1.2
Phase velocity 21.2.2
Phoria 4.6.4
Photography, infrared 22.5.2
Photographic instruments, looking down 18.4.3
Photographic instruments, looking up 18.4.2
Photomicrography 23.3.2.2
Photopic vision 4.4.4.9
Physical optics, restatement of principles 16.1.1.1
Physical structure of the eye 4.2.1
Physiological nystagmus 4.6.3
Pinhole interferometer, Young's 16.6
Pinhole size and contrast, discussion of 16.5.1
Plane of incidence 13.3.1
Plane-polarized light wave 3.2.2.1
Plane-polarized light wave, instantaneous magnitude 16.1.1.2
Plane, sagittal..................... 5.8.4.1
Planes, conjugate 6.5.7
Plastics, optical 17.6.2.3
Plate, diffraction 23.6.4
Point, far 7.2.5
Point, near 7.2.5
Point, virtual object 5.4.8.3
Points, cardinal 6.5.1.1
Polarizers 20.1.2.9
Polarization effects 20.7.1.2
Porro prism 13.7.6
Porro prism system 13.10.2
Porro prism system, Abbe's modification of........... 13.10.3
Porro prism tunnel.................. 13.8.2
Positive magnification 7.5.1.1
Power of the thin lens 6.7.2
Power, light gathering 18.3.3
Power, magnifying 6.3.7.2, 7.2.2
Power, resolving 26.1.2.6
Poynting vector, time-averaged 21.1.5.1
Presbyopia 4.4.4.2
Presbyopia 4.8.2
Pressure, intraocular 4.2.2
Principal planes, additional characteristics of 6.5.7
Principal plane, first 6.5.5.3
Principal plane, second 6.5.3
Prism, 45° - 90° - 45° 13.7.3
Prism, Abbe Type A 13.10.4
Prism, Abbe Type B 13.10.5
Prism, air-equivalent 13.8.3.1
Prism, Amici 13.10.7
Prism, Barr and Stroud ocular 13.10.19
Prism, Carl Zeiss ocular 13.10.18
Prism, double dove 13.10.11
Prism, Frankford Arsenal No.1 ..... 13.10.22
Prism, Frankford Arsenal No.2 ..... 13.10.23
Prism, Frankford Arsenal No.3 ..... 13.10.24
Prism, Frankford Arsenal No.4 ..... 13.10.25
Prism, Frankford Arsenal No.5 ..... 13.10.26
Prism, Frankford Arsenal No.6 ..... 13.10.27
Prism, Frankford Arsenal No.7 ..... 13.10.28
Prism, Harting-Dove 13.10.10
Prism, Leman 13.10.6
Prism, Pechan 13.10.12
Prism, penta 13.10.14
Prism, Porro 13.7.6
Prism, reversion 13.10.13
Prism, right angle 13.10.9
Prism, roof 13.7.4.2
Prism, Schmidt 13.10.8
Prisms, typical orientation of 13.9.1
Prism, Wallaston 13.10.15
Prism image rotation, 180 degrees 13.7.5
Prism length, reduced or apparent 13.8.3
Principal point, first 6.5.5.3
Principal point, second 6.5.3
Prism system, Carl Zeiss 13.10.16
Prism system, Carl Zeiss binocular-ocular 13.10.21
Prism system, Carl Zeiss coincidence.13.10.20
Prism system, Goerz 13.10.17
Prism system, Porro 13.10.2
Prism systems, telescope 13.7.4
Prism thickness, apparent 13.8.3.1
Prism thickness, reduced 13.8.3.1
Prism tunnel, Porro 13.8.2
Prism tunnel, right angle 13.8.1
Prisms and mirrors 13.7.1
Procedure, iteration 5.5.4.6
Procedure, refraction 5.3.1
Procedure, transfer 5.3.1
Profan 4.5.6.2
Protanomglous 4.5.6.1
Protanopes 4.5.6.1
Psycholigical and physical space variations, binocular 4.7.3
Pupil, entrance 6.11.3.1
Pupil, exit 6.11.3.2
Pupil, eye 4.4.2
Pupil, function 16.28.1.2
Pupil plane, entrance 6.11.3.1
Pupil plane, exit 6.11.3.2
Pupil point, entrance 6.11.3.1
Purkinje shift 4.5.2.1

Q-method 21.2.15
Quadrilayers 21.9
Quantities, first order 5.8.1.1
Quarternion methods 21.5
Quarternion's and corresponding matrices 21.5.2
Quarter wave bilayers 21.7.5
Quarter wave condition 21.3.5
Quarter wave miltilayers 21.10
Quarter wave stack 20.4.1, 20.4.3
Quarter wave stacks at non-normal incidence, reflectivity of 20.4.6
QWOT (Quarter wave optical thickness.) 20.1.3.4

Ramsden eyepiece 23.3.5.3, 14.4
Ratio, partial dispersion 2.7.3
Ray, axial 5.9.2.2
Ray, chief 6.3.1, 6.11.4, 8.3.1.1
Ray, close 5.8.1.1
Ray data, initial 5.4.6.1
Ray, differentially traced 5.8.1.1
Ray fan, meridional 8.3.1.1
Ray fan, skew 8.3.1.1
Ray, general 5.4.1.1
Ray, marginal 6.4.3.2

Ray, meridional ... 5.4.1.1
Ray, paraxial ... 5.4.1.1, 5.9.1
Ray, rim ... 6.4.3.2
Ray skew ... 5.4.1.1
Rays ... 2.2.1
Rays, direction of ... 2.1.3
Rays, types of ... 5.4.1
Ray trace, three element lens ... 6.2.5
Ray trace equations, paraxial ray ... 5.9.2
Ray trace equations, summary of ... 5.4.6
Ray trace equations, thin lens system in air ... 6.7.2
Ray trace format, mirror system ... 6.8.2
Ray trace format, single lens ... 7.2.4
Ray trace procedure, calculation of aberrations ... 6.2.6
Ray tracing ... 5.1.4.1
Ray tracing procedure, step by step ... 5.4.7
Real field ... 7.6.2.1
Recording optical tracking instrument ROTI Mark II ... 19.5.2
Reduced or apparent prism length ... 13.8.3
Reduced prism thickness ... 13.8.3.1
Reference indices ... 2.7.1
Refining method ... 20.4.8.3
Reflectance ... 17.2.2
Reflectance, energy ... 21.2.5.6
Reflectance and transmittance, methods of computing ... 20.1.5
Reflectance from plane parallel plates ... 16.16
Reflectances, complex, approximate method of computation ... 21.2.8.6
Reflecting darkfield condenser ... 23.4.3
Reflection coefficient ... 17.2.2
Reflection filters ... 20.1.2.10
Reflection, law of ... 2.3.2
Reflection, multiple ... 13.2.2.1
Reflection, surface ... 17.2.2
Reflective system, evolution of ... 19.3.1
Reflective system, Newton ... 19.3.2
Reflectivity ... 17.2.2
Reflectivity and transmittance ... 20.1.3.6
Reflectivity spectral ... 20.3.2.2
Reflector, heat ... 20.1.2.7
Reflectors, broad band ... 20.5.1.3.1
Reflectors, heat ... 20.6.2
Refracting darkfield condenser ... 23.4.2
Refracting material, characteristics of ... 17.2
Refraction equations ... 5.3.2
Refraction, index of ... 2.2.1
Refraction, law of ... 2.2.3
Refraction procedure ... 5.3.1
Refraction procedure at the aspheric surface ... 5.5.5
Refraction procedure at the spherical surface ... 5.4.5
Refractive index ... 3.2.1.2
Refractive indices, table of ... 2.5.3
Refractive indices, variation with wavelength ... 2.6.2.1
Refractive materials for specific wavelength ranges ... 17.6
Refractivity ... 2.7.2
Refractivity and dispersion, selection of material for ... 17.3.1
Relative index of refraction ... 2.5.2
Relay lens ... 7.3.2
Relay system, lens problem of a ... 12.2
Relay system, secondary color in a ... 12.4
Relay systems, double ... 12.6
Relay systems, lens ... 12.1
Relief, eye ... 7.3.8.2
Resolution, eye ... 4.4.5
Resolution, limit of ... 23.2.5.2
Resolution with circular apertures ... 16.27
Resolving power ... 26.1.2.6
Resolving power tests ... 26.1
Retardation, phase ... 21.6.2.2
Retina ... 4.4.4.5
Reversion prism ... 13.10.13
Revert ... 13.10.1.2
Right angle prism ... 13.10.9
Right angle prism tunnel ... 13.8.1
Right handed image, inverted ... 7.5.1.2
Right handed triplet solution ... 10.3.2.2
Rim ... 6.11.3.1
Rim ray ... 6.4.3.2
Ronchi test ... 25.9
Roof prism ... 13.7.4.2
Ross Baker system ... 19.4.3
Ross Baker system, modified ... 19.4.4

Sag or sagitta ... 5.5.2.2
Sagittal focus ... 5.8.4.1
Sagittal plane ... 5.8.4.1
Satellite tracking camera ... 19.5.1
Scattering, multiple ... 18.2.2.10
Schade system testing ... 26.3.1
Schmidt prism ... 13.10.8
Schmidt system ... 19.4.2
Scintillation ... 18.5.3
Scotopic vision ... 4.4.4.9
Secondary color ... 6.10.8.4
Secondary color in a relay system ... 12.4
Secondary spectrum ... 6.10.8.4, 19.2.2
Secondary spectrum in a doublet, reproduction of ... 11.4.2
Secondary spectrum in a triplet lens, correction of ... 11.4.3
Second focal length ... 6.5.2
Second focal plane ... 6.5.2
Second focal point ... 6.5.2
Second nodal point ... 6.5.4
Second principal plane ... 6.5.3
Second principal point ... 6.5.3
Second surface reflection ... 17.7.2
Seeing ... 4.1.2, 18.5.1.1
Seeing, atmosphere factors affecting ... 18.5.1
Seeing disc ... 18.5.3.2
Seidel longitudinal spherical aberration ... 8.5.3.2
Seidel theory of aberrations ... 19.4.2.7
Seidel tolerances ... 24.2.3
Seidel tolerances, use of ... 24.2.4
Semi-apochromatic microscope objective ... 23.3.4.1
Semi transparent mirrors ... 20.1.2.6, 20.8.2
Sensitivity, eye ... 4.5.1
Shadow bands ... 18.5.3.1
Shift, Purkinje ... 4.5.2.1
Shifted chief ray ... 8.9.2.1
Shifted chief ray, aberration polynomial ... 8.9.2
Short form ... 5.6.3.2
Short wave pass color filters ... 20.6.1
Short wave pass filters ... 20.6
Sign convention, angle ... 2.2.2

Simple magnifier, limitations of 7.3.1
Simple metal coatings 17.7.3
Simple microscope 7.3.2
Sine wave response test 26.4.4
Sine wave testing 26.2
Single layer coatings 20.3.3
Single lens 7.2.1
Single mirror imagery 13.4.1
Single wave, energy in a 3.2.3
Skew focus 5.8.4.1
Skew ray 5.4.1.1
Skew ray, initial data 5.4.2
Skew ray fan 8.3.1.1
Skew ray trace 8.1.2
Skew ray trace equations, use of 5.6.2
Slit interferometer, Young's 16.6.1.6
Slope angles 5.6.5.1
Small missile telecamera (SMT) 19.5.4
Smith-Helmholtz equations 6.3.8
Solutions, thin lenses 11.2.3
Spatial filtering 22.6.5.3
Spectra, channeled 16.19.1.1
Spectral reflectivity 20.3.2.2
Spectrum 2.6.1
Spectrum, secondary 6.10.8.4, 19.2.2
Spherical aberration 5.10.2.2
Spherical aberration coefficients 8.5.1.1
Spherical aberration, chromatic variation of 11.3.2.4
Spherical aberration, longitudinal 8.5.3.2
Spherical aberration, Seidel longitudinal 8.5.3.2
Spherical aberration, transverse 8.5.3.1
Spherical aberrations, zonal 8.5.2.4
Spherical darkfield condensers 23.4.5
Spherical wavefront 21.4.1
Spherical wavefronts, diffraction from 16.25
Sphere chromatism 19.2.5
Spheroid, oblate 19.4.2.6
Spot diagram 8.2.1
Square matrix 21.4.2.3
Stack, general P:G 20.4.4
Stack, multilayer 20.1.3.2
Stack, quarter wave 20.4.1, 20.4.3
Stacks with unequal thickness ratios 20.4.2
Staining, surface 2.7.5
Star test 25.11
Static visual acuity (SVA) 4.5.4.2
Stereoscopic depth 4.7.2.2
Steoeoscopy 4.7.2
Stiles Crawford effect 4.4.5.1
Stop shift equations 8.9.5.1, 9.3.4
Stop, aperture 6.11.2
Stop, field 7.3.1
Stop, funnel 23.4.3
Strabismus 4.6.4
Substage condenser 23.2.5.1
Substage condensers 23.3.3
Substrate, high index 20.3.4.5.1
Substrate, low index 20.3.4.3.1
Substrates for multi-layer filters 20.2.3
Summary of ray trace equations 5.4.6, 5.5.6
Summer heat haze 18.2.2.9
Superior oblique muscles 4.6.2
Superior rectus muscles 4.6.2
Surface coefficients, chromatic 6.10.3
Surface contributions 6.10.1
Surface contributions, third order spherical aberrations 8.5.2.1
Surface dimming 2.7.6
Surface reflection 17.2.2
Surface staining 2.7.5
System data, initial 5.4.6.1
System, double half wave 20.10.7
System, telescope 6.3.7.1

Table of critical indices and angles 2.4.2
Table of optical glass characteristics 2.7.8
Table of refractive indices 2.5.3
Talbot's law 4.5.3
Tangential focus 5.8.4.3
Target detection and location, infrared 22.6.5
Target, USAF resolution 26.1.3
Target value 5.5.7.2
Taylor triplet lens 10.1.1
Telescopes, angular aberrations of 11.1.3.5
Telescope, completed design 15.5
Telescope, Galilean 7.6
Telescope design, typical example of 15.2.4
Telescope , doublet 11.2.1.1
Telescope objectives, secondary spectrum of 11.4
Telescope objective and eyepiece design 7.4.3
Telescope objective system, Petzval curvature of 11.1.3
Telescope magnifying power 7.4.2
Telescope prism systems 13.7.4
Telescope system, Cassagrain 19.3.3
Telescopic system 6.3.7.1
Terms, deformation 5.5.23
Test, astigmatism and curvature of field 25.5
Test, Coltman variable frequency square wave 26.4.2
Test, distortion 25.6
Test, fixed frequency square wave 26.4.3
Test, focal length 25.2
Test, Foucault 25.10
Test, Jentsch's grid method 25.9.4
Test, longitudinal special aberration 25.3
Test, NBS resolving power 26.1.2
Test, Ronchi 25.9
Test, resolving power 26.1
Test, sine wave response 26.4.4
Test, star 25.11
Tester, Ann Arbor 25.9.3
Testing, design phase 24.1.4.2
Testing, evaluation phase 24.1.4.4
Testing, manufacturing phase 24.1.4.3
Testing, Lambert system 26.3.2
Testing, Schade system 26.3.1
Testing, sine wave 26.7
Theory, waves 3.2.1
Thermal effects, types 18.6.1
Thick lens doublet 11.3.1
Thick lens telescope objective design procedure 11.3
Thickness, lens 10.4.1
Thin films 17.7.1
Thin films, general uses of 20.1.1
Thin films, properties of 21.1.2
Thin films, uses of 21.1.1
Thin film coatings, typical application of 20.1.2

Thin film materials ... 20.2.4
Thin lens ... 6.7.1.2
Thin lens, chromatic aberration in ... 6.10.7.1
Thin lens, first order ... 11.2.1
Thin lens, power of ... 6.7.2
Thin lens, third order ... 11.2.2
Thin lens aberration, third order coefficients ... 8.10.1
Thin lens achromatic system ... 6.10.8
Thin lens concept ... 6.7.1
Thin lens solutions ... 11.2.3
Thin lens telescope objective, design procedure of ... 11.2
Third order, thin lens ... 11.2.2
Third order aberration coefficient ... 8.5.2
Third order aberration coefficients, value of ... 8.7.4
Third order aberrations, adjusting ... 9.2.4.10
Third order aberrations, examples of ... 8.6.2
Third order aberrations of doublet objective lens, automatic correction of ... 11.3.2
Third order aberrations of prisms ... 13.9.2
Third order coefficients, evaluation of ... 10.3.1
Third order contributions, basic formulae ... 8.7.1
Third order optics ... 5.11.3.1
Third order polynomial, afocal ... 8.8.1
Third order spherical aberration surface contributions ... 8.5.2.1
Third order surface contributions ... 9.3.3
Three element lens, ray trace
Time averaged energy density ... 3.1.2.2
Time averaged energy density ... 16.1.1.3
Time averaged Poynting vector ... 21.2.5.1
Tolerances, Seidel ... 24.2.3
Total chromatic aberration ... 6.10.3
Trace, skew ray ... 5.4.1.3
Traced ray, diferrentially ... 5.8.1.1
Tracing, ray ... 5.1.4.1
Transfer equations ... 5.3.1
Transfer procedure ... 5.3.1
Transfer procedure, physical surface to next tangent plane ... 5.4.3
Transfer procedure, tangent plane to aspheric surface ... 5.5.4
Transfer procedure, tangent plane to spherical surface ... 5.4.4
Transmission of light ... 17.2.1
Transmittance, complex ... 21.2.5.6
Transmittance of plane parallel plates ... 16.15
Transmittance, internal ... 17.2.3.2
Transparency losses contributing to extinction, types of ... 18.2.1
Transverse axial chromatic aberration (TAch) ... 6.10.3
Transverse chromatic aberration ... 6.10.1
Transverse chromatic aberration (Tch) ... 6.10.3
Transverse displacement ... 6.10.3
Transverse spherical aberration ... 8.5.3.1
Triple layer antireflection coatings ... 20.3.5
Triggered radiation type infrared system ... 22.5.4
Trilayers ... 21.8
Triplet lens corrected for secondary color ... 11.4.5
Triton ... 4.5.6.2
Trophias ... 4.6.4
True angular field, microscope ... 23.2.3
Tube, image converter ... 22.2.1
Tube currents ... 18.6.2
Tube length, optical ... 23.2.3
Tunnel diagram ... 13.8.1.1
Two mirror image, invariant position of ... 13.6.3
Two mirror image location ... 13.6.1
Two mirror image location axiom ... 13.6.2
Twyman-Green interferometer ... 25.8.4.4
Twyman-Green interferometer, principles of operation ... 16.3.1
Type A vertical illuminator ... 23.3.2.4.1
Type B vertical illuminator ... 23.3.2.4.2
Type of coating, choice of ... 20.3.2
Types of rays ... 5.4.1
Typical orientation of prisms ... 13.9.1

Uncoated surface, reflectivity of ... 20.3.1
Ultraviolet ... 2.6.1
Undercorrected lens ... 6.10.5
Unit positive magnification ... 6.5.7
U.S.A.F. resolution target ... 26.1.3
Uses of mirrors and prisms ... 13.1.1

Value, nu ... 2.7.2
Value, target ... 5.5.7.5
Variation of illumination ... 25.10.2.2
Variations of refractive indices with wavelength ... 2.6.2.1
Vector form of the law of refraction ... 2.2.4
Vector ray tracing matrix form ... 13.4.3
Vector, Poynting, time averaged ... 21.2.5.1
Vectors, electric and magnetic ... 3.2.1.1
Vee number ... 2.7.2
Velocity of electromagnetic waves ... 3.2.1.2
Velocity, phase ... 3.2.1.2, 21.2.2
Vertical illumination ... 23.2.4
Vertical, illuminator, type A ... 23.3.2.4.1
Vertical illuminator, type B ... 23.3.2.4.2
Vertical microscope illuminator ... 23.3.2.4
Vibration, period of ... 3.2.1.2
View, field of ... 7.3.1, 22.6.3.2
Vignetting ... 6.11.8.1
Virtual object point ... 5.4.8.3
Vision, color ... 4.5.6
Vision, dichromatic ... 4.13
Vision, loss of ... 4.1.3
Vision, measuring ... 4.5.4
Vision, mesoptic ... 4.4.4.9
Vision, photopic ... 4.4.4.9
Vision, scotopic ... 4.4.4.9
Visual acuity, (VA) ... 4.5.4
Visual acuity, dynamic (DVA) ... 4.5.4.2
Visual acuity, static (SVA) ... 4.5.4.2
Visual angles ... 4.4.4.5
Visual instruments, imposed limitations ... 18.3.1
Visual system, numerical example ... 12.3
Vitreous humor ... 4.4.4.4

Wavefront, spherical ... 24.4.1
Wavefront, cutoff ... 20.5.1.2
Wavelength range, infrared ... 22.4.3
Wave number ... 20.1.3.3
Wave, surfaces and rays ... 2.1.2

Waves collinear, coherent 3.3.1
Waves collinear, incoherent 3.3.2
Waves, non-collinear, coherent 3.3.3
Wave theory 3.2.1
Width, fringe 3.3.3.3
Window 26.1.3.2
Wollaston prism 13.10.15

Young's pinhole interferometer 16.6
Young's slit interferometer 16.6.1.6

Zero condition 21.2.16
Zeroth approximation 5.11.1.4
Zeroth order 5.11.1.4
Zone, high reflectance 20.4.1.2
Zone aberration,
  methods of reducing 11.3.4
Zonal aberration, tolerance on 11.3.3
Zonal aberration correction
  methods, discussion of 11.3.5
Zonal numerical apertures 16.26.1.3
Zonal spherical aberration 19.2.4

☆ U. S. GOVERNMENT PRINTING OFFICE: 1963 O - 709-648